Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction

Marianna Apidianaki

▶ **To cite this version:**

Marianna Apidianaki. Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction. Linguistique. Université Paris-Diderot - Paris VII, 2008. Français. NNT : . tel-00322285

HAL Id: tel-00322285
https://theses.hal.science/tel-00322285

Submitted on 17 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PARIS. DIDEROT (Paris 7)

ECOLE DOCTORALE : Sciences du Langage
Laboratoire LATTICE – CNRS UMR 8094

DOCTORAT
Linguistique Théorique, Descriptive et Automatique

MARIANNA APIDIANAKI

Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction

Thèse dirigée par Catherine Fuchs

Soutenue le 5 septembre 2008

JURY

M. Helge DYVIK	Professeur, Université de Bergen
Mme Catherine FUCHS	Directeur de recherche, CNRS (Directeur)
M. Eric GAUSSIER	Professeur, Université Joseph Fourier, Grenoble I
M. Daniel KAYSER	Professeur, Université Paris Nord (Rapporteur)
M. Philippe LANGLAIS	Professeur, Université de Montréal (Rapporteur)
Mme Elsa SKLAVOUNOU	Technical Account Manager SYSTRAN

À mes parents

REMERCIEMENTS

En arrivant à la fin de cette thèse, j'aimerais remercier ceux qui m'ont soutenue tout au long de ce parcours.

Tout d'abord, je tiens à remercier ma directrice, Catherine FUCHS, de m'avoir accueillie au sein du LATTICE et de m'avoir orientée dans mes recherches pendant mes années de thèse. Je tiens également à remercier Laurence DANLOS, actuelle directrice du LATTICE, de son soutien.

J'adresse mes plus vifs remerciements aux professeurs Daniel KAYSER et Philippe LANGLAIS, pour m'avoir fait l'honneur d'accepter d'évaluer ce travail. Leurs commentaires extrêmement rigoureux, précis et par conséquent précieux m'ont été extrêmement profitables. Pour avoir accepté de participer à ce jury de thèse et d'examiner mon travail, je tiens tout particulièrement à remercier les professeurs Helge DYVIK et Eric GAUSSIER et Mme Elsa SKLAVOUNOU.

Je tiens aussi à exprimer ma gratitude envers les membres du laboratoire RALI (Université de Montréal) et à Didier BOURIGAULT (Université de Toulouse-Le Mirail) pour le support qu'ils m'ont apporté.

Ce travail de thèse a été financé par l'Institut de Bourses de l'Etat hellénique (IKY). Il a été également subventionné en partie par une Bourse d'accueil Marie Curie pour la formation de chercheurs en début de carrière (Early Stage Research Training Fellowship), dans le cadre du projet MULTILINGUA à l'Université de Bergen (Norvège). Je remercie vivement les professeurs Helge DYVIK et Koenraad DE SMEDT de m'avoir donné la possibilité de profiter de cette expérience scientifiquement enrichissante, ainsi que le personnel du centre Aksis pour son accueil chaleureux.

Ce que cette thèse représente réellement pour moi, peu de gens le savent. A ces amis, merci de m'avoir accompagnée durant ce travail et d'avoir été si présents, et ce, parfois malgré la distance. Je ne pourrais pas clore ce passage sans adresser mes plus vifs remerciements à Sandrine et à Djamé pour leur soutien précieux lors de la dernière ligne droite.

To translate is one thing;
to say how we do it, is another.

Haas W. (1962), Philosophy, vol. 37

Le sens n'a de sens que
si on lui en donne et vice versa

Anonyme, fin du XX^{e} siècle
(cité par Kleiber, 1999)

SOMMAIRE

INTRODUCTION

1. Avant propos

L'**ambiguïté lexicale** est un phénomène omniprésent dans les langues naturelles, qui en constitue une des sources de richesse et de souplesse. Dans le cadre de la communication humaine, l'ambiguïté ne constitue pas un réel problème en raison des inférences effectuées par les participants dans l'acte de communication. Ces inférences s'appuient aussi bien sur les informations linguistiques et extra-linguistiques accessibles aux agents lors de la production et de la réception d'un énoncé, que sur leurs connaissances du monde, mobilisées par leur volonté de communiquer.

Néanmoins, dans un cadre de **traitement automatique de la langue**, l'ambiguïté lexicale constitue toujours un **défi**. La difficulté à traiter automatiquement ce phénomène a été reconnue très tôt, dès l'apparition, en fait, des premiers systèmes de Traduction Automatique (Bar-Hillel, 1960). Cette difficulté a même constitué l'une des raisons de douter des potentialités de l'application en question. La complexité à modéliser et à traiter l'ambiguïté lexicale a ainsi provoqué sa marginalisation, favorisant le développement d'applications focalisées sur des domaines restreints, bien délimités, sur la base de l'hypothèse que, précisément, le problème de l'ambiguïté ne s'y pose pas. Ce type de limitation imposant, à son tour, des limites à l'extensibilité des systèmes, l'intérêt pour le traitement de l'ambiguïté lexicale s'est de nouveau manifesté ; ainsi, l'analyse, la modélisation et le traitement de ce phénomène constituent de nos jours des champs d'étude faisant l'objet de nombreux travaux.

2. Problématique

La notion de **sens** peut être conçue de diverses manières, selon le cadre théorique adopté. Outre cette diversité, les principes sous-jacents à une modélisation de la sémantique lexicale sont fortement influencés par les besoins auxquels une telle modélisation est supposée répondre. Dans un **cadre automatique**, par exemple, des inventaires sémantiques sont requis pour la **désambiguïsation lexicale** dans des applications précises. Au sein de ces applications, la désambiguïsation ne constitue qu'une tâche intermédiaire de traitement, qui facilite leur objectif final. Ainsi, les besoins des applications en matière de désambiguïsation divergent et sont définis par leur finalité.

Cependant, la plupart des méthodes de désambiguïsation existantes sont indépendantes d'un cadre applicatif précis. Et servent, souvent, à éprouver la validité de la théorie de traitement linguistique sur laquelle elles se fondent, sans prendre réellement en considération les enjeux d'ordre pratique. Ainsi, tout en considérant cette tâche intermédiaire de traitement comme importante pour pouvoir atteindre un objectif final, l'évaluation de ses résultats ne tient, bien souvent, pas compte de cette finalité. L'incohérence constatée entre le statut de la tâche de désambiguïsation et le cadre dans lequel les méthodes sont développées et évaluées a comme résultat la non-conformité des méthodes à des applications précises. Ce constat a suscité l'intérêt pour des méthodes de désambiguïsation adaptées aux besoins des applications, et dont l'évaluation se fait par référence aux résultats finaux.

Outre les méthodes employées, ce besoin de conformité à la finalité des applications concerne également les inventaires de sens utilisés pour la désambiguïsation. Des inventaires construits pour des usagers humains sont, en effet, inadaptés à la désambiguïsation dans des applications automatiques et même les inventaires développés pour être exploités dans le cadre de certaines applications ne s'avèrent pas conformes à d'autres. C'est le cas, par exemple, de l'exploitation dans un cadre bilingue d'inventaires développés dans un cadre monolingue, où les distinctions sémantiques proposées ne correspondent pas aux besoins d'un traitement bilingue.

La nécessité d'élaborer des **méthodes de désambiguïsation lexicale orientées vers des applications** précises s'est donc, récemment et progressivement, imposée et de telles méthodes ont désormais vu le jour. En ce qui concerne la construction d'inventaires sémantiques, l'inadaptation des ressources prédéfinies à un traitement automatique, d'un côté, et les grandes divergences constatées au niveau des descriptions issues de ressources différentes, d'un autre côté, ont provoqué l'émergence de méthodes automatiques, permettant de produire de telles ressources à partir de données textuelles.

Notre travail s'inscrit dans le courant des études visant une analyse sémantique et une modélisation orientées vers un type d'applications précis. Dans notre cas, il s'agit des **applications de traduction**, à savoir la **Traduction Automatique** et la **Traduction Assistée par Ordinateur**. L'analyse de la sémantique lexicale effectuée se situe donc dans une perspective de modélisation conforme au traitement dans un contexte bilingue. Cette modélisation définit des correspondances inter-langues à un niveau plus élevé que le niveau lexical, celui du sens.

Hormis la question de la conformité à un cadre précis, l'analyse sémantique effectuée permet d'aborder d'autres questions, qui ont également un fort impact sur la modélisation. Ainsi la **distinctivité** des sens lexicaux et leur **énumération**. La difficulté à proposer une réponse unique quant au nombre et à la granularité des sens se reflète dans les divergences observées entre ressources différentes, et qui constituent d'ailleurs la raison principale à leur incompatibilité.

La majorité des ressources existantes partagent, en outre, la même incapacité à représenter les **liens inter-sens** et, par conséquent, à **différencier** les sens par rapport à leur **statut**. Ce manque de flexibilité ne peut se justifier dans un cadre de sémantique lexicale, où nous pouvons davantage parler de régions sémantiques au sein desquelles des glissements et des altérations ont lieu, que d'un cadre où des délimitations claires et nettes peuvent s'appliquer. L'absence de différenciation entre sens relativement à leur statut au sein des inventaires produit un traitement uniforme de sens de nature différente. Elle influence également l'évaluation des méthodes qui se fondent sur ce type d'inventaires. Ainsi, lors de l'évaluation des processus de désambiguïsation, par exemple, une

même pénalisation s'applique à des erreurs plus ou moins importantes, impliquant des sens proches ou distants.

La modélisation sémantique que nous proposons permet, quant à elle, de prendre en compte la **nature** et les **liens** des **sens lexicaux**. Grâce à ces liens, les représentations sémantiques engendrées rendent possible une description à **granularité variable**, exploitable conformément aux besoins qui se présentent. Cette ressource sémantique est élaborée à l'aide de notre **méthode d'acquisition de sens**.

Son caractère inter-langue, sa structure ainsi que la nature des sens proposés, la rendent particulièrement apte à être utilisée comme inventaire de sens pour la **désambiguïsation lexicale** qui s'opère dans des applications de traduction. Notons, néanmoins, qu'une pratique couramment répandue dans ce cadre consiste à définir des **correspondances biunivoques** entre sens des mots source et équivalents de traduction. Cette hypothèse attribue à chaque équivalent le même rôle en tant qu'**indice de sens**, ce qui empêche une différenciation des sens induits relativement à leur statut. Notre modélisation répond à cette question en différenciant le rôle des équivalents en tant qu'indices de sens ; ainsi, certains induisent des sens clairement distants, tandis que d'autres servent à repérer des distinctions sémantiques moins nettes, et ce, en étant souvent regroupés avec d'autres équivalents des mots source.

Etant donné la non bi-univocité de nos correspondances inter-langues, les résultats de la **méthode de désambiguïsation** que nous avons mise en place se situent donc à un niveau plus abstrait que le niveau lexical : le niveau sémantique, qui ne concerne pas toujours des mots isolés. Afin de pouvoir atteindre le niveau lexical (finalité des applications de traduction), un besoin de filtrage s'impose. Ce filtrage, qui consiste précisément en une **sélection lexicale**, s'opère sur les résultats de la désambiguïsation et permet d'émettre des propositions de traduction sémantiquement pertinentes pour les instances des mots ambigus source. Les résultats du filtrage effectué sont ensuite évalués par une métrique s'appuyant, elle aussi, sur les informations contenues dans la ressource sémantique construite et rendant la prise en compte de la pertinence sémantique des prédictions de traduction possible.

3. Plan de la thèse

Notre travail débute par une réflexion théorique sur le phénomène de l'**ambiguïté lexicale** et sur les facteurs influençant son analyse et sa modélisation, à la fois dans un cadre monolingue et dans un cadre bilingue. Nous insistons essentiellement sur les particularités d'une telle analyse au sein d'un **cadre bilingue**, particularités qui découlent du besoin de mettre en correspondance des unités lexicales de langues différentes se trouvant en relation de traduction.

Par l'analyse des facteurs provoquant les divergences observées au sein d'inventaires sémantiques différents, nous procédons, dans un deuxième temps, à une présentation de **méthodes d'acquisition d'informations sémantiques** à partir de données textuelles. Il s'agit, d'une part, de méthodes fondées sur des techniques d'apprentissage non supervisé, principalement utilisées dans un cadre monolingue et, d'autre part, de méthodes traductionnelles, s'appuyant sur les informations issues de corpus parallèles.

Nous présentons ensuite un ensemble de **méthodes de désambiguïsation lexicale** basées sur des connaissances et dirigées par les données. Nous exposons les spécificités de ces deux types de méthodes et expliquons leurs avantages et leurs inconvénients relativement au traitement automatique. Quel que soit le type de méthode utilisée, la désambiguïsation de nouvelles instances de mots ambigus passe toujours par l'exploitation d'informations provenant de leurs contextes ; une grande partie de l'analyse est donc dédiée à la notion de **contexte**, dans un cadre monolingue et bilingue, et aux différentes manières possibles de prendre en compte les informations qui le constituent. En outre, la question de l'évaluation des méthodes de repérage de sens et de désambiguïsation fait l'objet d'une analyse détaillée, qui montre les difficultés surgissant lors de cette tâche ainsi que les tentatives faites pour sa standardisation.

Nous présentons, dans un troisième temps, les **méthodes** développées dans le cadre de ce travail de thèse. Toutes les méthodes que nous proposons sont **dirigées par des données** extraites d'un corpus bilingue parallèle. Les caractéristiques du corpus utilisé pour l'apprentissage et les différentes étapes de prétraitement qu'il a subies sont donc, tout d'abord, exposées. Nous procédons ensuite à la présentation d'une **méthode monolingue d'acquisition de sens**, que

nous avons initialement développée dans le but d'établir des correspondances sémantiques inter-langues. Puis, nous démontrons les inconvénients de l'utilisation d'une méthode de ce type dans un cadre de traduction, inconvénients sur lesquels nous nous sommes appuyée pour définir une autre méthode de repérage de sens, plus conforme au traitement dans des applications de traduction.

Il s'agit d'une méthode d'**apprentissage automatique non supervisé** opératoire dans un cadre bilingue, qui s'appuie sur des informations distributionnelles et traductionnelles extraites du corpus parallèle d'apprentissage. Contrairement à la première méthode utilisée, les informations de traduction entrent, ici, en jeu dès le début. L'analyse sémantique effectuée permet de repérer les sens des mots ambigus de la langue source en s'appuyant sur la clusterisation de leurs équivalents dans la langue cible.

Nous présentons les principes de fonctionnement de cette méthode ainsi que les détails de l'implémentation. Nous insistons sur le processus de **calcul de similarité** entre les équivalents et sur la **technique de clustering** utilisée. Ensuite, nous illustrons les résultats de la méthode d'acquisition de sens à l'aide d'exemples précis et concluons par une présentation des points forts et des points faibles de la méthode proposée. La notion de **similarité sémantique**, notion centrale de la méthode de repérage de sens et qui joue aussi un rôle important lors de l'exploitation des résultats pour la traduction – dans la mesure où elle définit la commutabilité des équivalents de traduction similaires des mots ambigus – est analysée par la suite.

Notre **méthode de désambiguïsation lexicale**, fondée sur l'exploitation des résultats de la méthode d'acquisition de sens, est alors présentée. Cette méthode entreprend de sélectionner, parmi les sens décrits au sein de l'inventaire sémantique généré, le sens véhiculé par de nouvelles instances des mots ambigus dans un corpus de test. Les principes et le fonctionnement de notre **méthode de sélection lexicale**, mise en place dans le but de raffiner les résultats de la désambiguïsation pour la traduction, sont alors présentés.

Nous concluons par une présentation des étapes d'évaluation que nous avons menées : une **évaluation qualitative** des résultats de la méthode d'acquisition de sens et une évaluation quantitative de la pertinence des résultats

des méthodes de désambiguïsation et de sélection lexicale. Nous retrouvons, sur ce point également, l'influence de notre modélisation sémantique, dont les particularités ont permis l'élaboration d'une métrique pondérée du résultat de la sélection lexicale, caractérisée par sa capacité à valoriser des propositions de traduction sémantiquement pertinentes. Le fonctionnement de la métrique en question est donc exposé dans le dernier chapitre, ainsi que les résultats issus de cette évaluation.

CHAPITRE 1

L'AMBIGUÏTE LEXICALE

INTRODUCTION

Le terme **ambiguïté lexicale** décrit de nombreux phénomènes. Dans son acception la plus générale, il désigne des cas d'unités lexicales caractérisées par un certain nombre de distinctions sémantiques, sans précisions supplémentaires quant à la nature de ces distinctions. Or, employer ce terme au sens large sous-entend une uniformité au niveau du traitement des différents cas d'ambiguïté, au lieu d'une différenciation en fonction des besoins. Un autre corollaire de cette absence de précision est la généralisation de l'emploi du terme de **désambiguïsation lexicale** qui sert à décrire la résolution de différents types d'ambiguïté.

Dans ce chapitre, nous allons, dans un premier temps, analyser un ensemble de questions liées au repérage des sens lexicaux et à leur nature, et examiner la possibilité d'établir une typologie des unités lexicales ambiguës. Dans un second temps, nous allons regarder la manière dont l'ambiguïté lexicale peut être prise

en compte dans un cadre bilingue et les relations de traduction qu'elles peuvent entretenir avec des unités lexicales d'autres langues. Ce chapitre introductif nous permettra de clarifier certaines notions nécessaires pour la suite de cet exposé et d'analyser, d'un point de vue théorique, certains phénomènes que nous avons rencontrés.

1. Repérage des sens de mots ambigus

1.1. Délimitation des sens lexicaux

La délimitation des sens des mots est une question à laquelle il est difficile de trouver une solution unique et universellement acceptable. Sur un plan théorique, les réponses qui peuvent être données sont souvent aussi nombreuses que les approches adoptées au sein des différents courants de sémantique lexicale. Sur un plan pratique, la complexité de la tâche se reflète dans les divergences observées tant au niveau des descriptions des sens données dans des ressources lexicales différentes qu'à celui des descriptions fournies par des méthodes automatiques de repérage de sens. La variabilité des descriptions proposées montre la difficulté de la tâche d'identification du sens et souligne qu'un consensus sur le nombre de sens des unités lexicales et leur granularité est loin d'être atteint.

La variation observée au niveau des distinctions sémantiques peut être attribuée à un grand nombre de facteurs, dont celui de l'absence d'accord général sur la nature du sens lexical. L'une des sources de ce désaccord repose sur la possibilité d'appréhender la question du sens sous plusieurs « angles » : en tant que « sens référentiel » dans le cadre de la **sémantique véric onditionnelle** [1] ; en termes de rapport entre signes et représentations mentales dans le cadre de la **sémantique psychologique**[2] ou de la **sémantique cognitive**[3] ; de manière intra-

[1] La **sémantique véric onditionnelle**, développée dans le cadre des théories **référentielles** ou **dénotationnelles**, considère le sens d'une expression en termes de conditions de *vérité* (Frege, 1892 ; Tarski, 1944) ou de *satisfaction* qui doivent être remplies pour que la référence à des occurrences particulières au moyen de l'expression puisse avoir lieu (Kleiber, 1999 : 32). On parle, dans ce cas, de **sens référentiel**.

[2] Sémantique développée dans le cadre de la psycholinguistique qui considère la signification comme le rapport entre signes et représentations ou opérations mentales.

linguistique dans le cadre de la **sémantique différentielle**[4] ; en tant que produit des échanges linguistiques sur la base d'un ensemble d'éléments plus ou moins stables (Victorri, 1997 : 53), ou d'une manière contextualisée, qui le rapproche de la notion d'usage des mots dans les textes (Wittgenstein, 1953 ; Firth, 1957a,b ; Harris, 1954). Cette dernière approche est la plus souvent adoptée dans la pratique lexicographique d'aujourd'hui, et fait correspondre le sens des unités lexicales à l'ensemble de leurs différents usages.

La diversité des conceptions quant à la nature du sens rend difficiles tant la proposition que l'acceptation d'une réponse unique sur le nombre et la granularité des distinctions sémantiques qui caractérisent les unités lexicales et explique ainsi, en partie, la complexité impliquée dans les tentatives d'identification et de description de ces unités. Des considérations de ce type ont aussi une influence sur la conception des ressources sémantiques et expliquent, par conséquent, et jusqu'à un certain point, les divergences présentes au niveau des distinctions sémantiques entre ressources différentes.

Toujours sur le plan théorique, et liée à la nature des sens lexicaux, se pose la question de leur **distinctivité**. La possibilité d'établir des distinctions claires entre sens dépend de la manière dont ils se différencient entre eux, c'est-à-dire de la relation qu'ils entretiennent (ou non) entre eux. L'estimation de la distinctivité des sens présente un grand intérêt théorique, étant donné qu'elle peut servir à la définition de degrés d'ambiguïté et à la discrimination entre types différents d'unités ambiguës. Néanmoins, la question même de la typologie constitue un champ de controverses important. Non seulement les catégories proposées dans la littérature varient beaucoup d'un courant à l'autre, mais la possibilité même d'établir des catégories est parfois mise en cause. Inhérente au sujet de la typologie, nous retrouvons, évidemment, la question des propriétés sémantiques

[3] Considérée comme un développement de la sémantique psychologique, la sémantique cognitive considère également le sens comme une représentation mentale. Une des théories développées dans ce courant est la **sémantique du prototype** (Rosch, 1973 ; Kleiber, 1990), selon laquelle le sens dérive d'un certain degré de proximité avec un prototype, un exemplaire modèle pour une catégorie.

[4] Issue de la linguistique structurale, selon laquelle le sens est intra-linguistique et se construit par des relations d'opposition entre les unités lexicales du système linguistique (Greimas, 1986). Dans une telle optique, le sens d'un mot est défini par les différences qu'il entretient avec ceux des autres mots du vocabulaire de la langue, différences représentables par des **sèmes** ou des **traits sémantiques**.

des mots, en fonction desquelles leur attribution à tel ou tel type d'ambiguïté serait effectuée.

Outre son intérêt théorique évident, le traitement de ces questions présente également un grand intérêt pratique, dans la mesure où il rendrait possible la définition de critères de description de la sémantique des mots dans la pratique lexicographique. En effet, les variations dans la conception des types d'ambiguïté et des caractéristiques des unités attribuées à chaque type, se reflètent dans les divergences observées entre les ressources, tant au niveau du nombre des entrées qu'à celui de leur contenu (c'est-à-dire, la répartition des sens à l'intérieur des entrées). Mais avant d'aborder la description de ces divergences, nous allons analyser les principales propositions de catégorisation des mots ambigus.

Les niveaux d'ambiguïté les plus fréquemment traités dans la littérature sont l'**homonymie** et la **polysémie**. La différence principale entre ces deux niveaux repose sur l'existence d'un **lien** entre les sens du mot polysémique. Ainsi, la polysémie décrit les cas où, à un signifiant unique, correspondent plusieurs sens perçus comme liés entre eux, tandis qu'à l'opposé, l'homonymie décrit des cas caractérisés par une absence de lien entre les sens identifiés (Martinet, 1974 ; Lyons, 1978[5]). Plus précisément, dans le cas de l'homonymie, nous avons à faire à deux signes distincts. Intéressante est la position de Polguère sur la question (2003 : 126-127). Selon Polguère, l'homonymie peut être perçue, non pas comme une vraie relation sémantique entre lexies, mais plutôt comme une relation de forme très forte, une identité de signifiants, qui est particulière en ce qu'elle s'accompagne précisément d'absence de lien sémantique[6].

[5] Traduit dans *Sémantique linguistique*, p. 178-196, Larousse, Paris, 1990, traduction française de l'édition anglaise, Cambridge University Press, 1978.

[6] D'après Polguère, les homonymes sont deux lexies « *qui sont associées aux même signifiants, mais ne possèdent aucune intersection de sens notable* » ou « *qui (…) n'entretiennent aucune relation de sens* » (2003 : 126-127). Il s'agit donc d'une absence de relation sémantique, perçue comme remarquable en ce qu'elle contraste avec la présence d'une identité de forme. En revanche, la polysémie est, pour Polguère, la propriété d'un « vocable » de contenir plus d'une lexie. Dans la terminologie de Polguère, une « lexie » est un regroupement de mots-formes (lexèmes) ou de constructions linguistiques que seule distingue la flexion (locutions) et chaque lexie est associée à un sens donné, que l'on retrouve dans le signifié de chacun des signes (mots-formes ou constructions linguistiques) auxquels elle correspond (*ibid.* : 50-51). En revanche, un « vocable » est un regroupement de lexies associées aux même signifiants et qui présentent un lien sémantique évident (par ex. *verre* : « type de contenant » – « matériau »). Les lexies d'un vocable sont souvent appelées ses acceptions. Une lexie est alors un regroupement de signes et un vocable, un regroupement de lexies.

Cette distinction « classique » entre homonymie et polysémie se retrouve sous d'autres dénominations. Par exemple, chez Weinreich (1964) le terme **ambiguïté contrastive** décrit les cas où un élément lexical porte deux sens distincts et non liés, tandis que le terme **polysémie complémentaire** concerne les cas où les sens lexicaux sont des manifestations du même sens principal du mot, quand il apparaît dans des contextes différents. Il existe pourtant une différence entre ambiguïté contrastive et homonymie, en ce sens que dans le premier cas, un seul élément lexical est impliqué, tandis que dans le deuxième, l'existence de deux éléments lexicaux distincts est sous-entendue.

Un paramètre, qui serait de grande utilité s'il était pris en compte pour distinguer homonymie et polysémie, concerne les **propriétés grammaticales** des mots[7]. Ces propriétés, habituellement ignorées, permettraient d'affiner et d'enrichir la distinction décrite ci-dessus. Pustejovsky (1995 : 28), quant à lui, analyse ce type de propriétés et décrit par le terme de **polysémie complémentaire** la relation entre les sens de catégories différentes du mot polysémique. Les cas d'ambiguïté complémentaire qui n'impliquent aucun changement de catégorie lexicale et dans lesquels les différents sens du mot se recouvrent, sont interdépendants ou possèdent des éléments sémantiques en commun, relèvent, selon Pustejovsky, de la **polysémie logique** [8]. Cette prise en compte des propriétés grammaticales des mots se retrouve chez Hirst (1987 : 5-6), qui distingue à son tour trois types d'ambiguïté lexicale : la **polysémie**, **l'homonymie** et l'**ambiguïté catégorielle** [9].

[7] Dans notre étude, nous nous intéressons à la polysémie des mots au sein d'une catégorie grammaticale ; nous n'allons donc pas traiter les cas de polysémie complémentaire (ou ambiguïté catégorielle). Nous considérons d'ailleurs que ce type d'ambiguïté peut souvent être résolu par l'étiquetage morphosyntaxique des textes (Wilks et Stevenson, 1997). La prise en compte des parties du discours agit effectivement comme un filtre, en réduisant le nombre de sens candidats d'un mot cible (Vasilescu *et al*., 2004).

[8] Pustejovsky décrit comme cas de polysémie logique, les cas d'altérations sémantiques où le mot semble avoir des sens liés de manière systématique. Telles sont, par exemple, les altérations : dénombrable/massif, récipient/contenu, produit/producteur, processus/résultat, plante/nourriture, endroit/personnes (*ibid*., p.31). Nous reviendrons sur cette notion un peu plus loin.

[9] La distinction entre les deux premiers types se définit par la présence (par ex. les sens de *open* : « unfolding », « expanding », « reveling », « moving to an open position », etc.) ou l'absence de lien entre sens, tandis que les mots ambigus du point de vue de la catégorie sont ceux dont la catégorie syntaxique peut varier. L'ambiguïté catégorielle est illustrée par le mot *sink*, qui peut être un substantif décrivant une « plumbing fixture » ou un verbe signifiant « become submerged ». Hirst procède aussi à la distinction de sous-catégories au sein de l'homonymie : il parle d'**homographes** (dans le langage écrit), mots dont les sens sont associés à un seul lexème bien que peuvant être

Les types d'ambiguïté lexicale décrits ci-dessus peuvent caractériser un mot conjointement. Ainsi, il est possible qu'un mot soit simultanément affecté par l'homonymie et la polysémie, par l'homonymie et l'ambiguïté catégorielle, ou encore par la polysémie et l'ambiguïté catégorielle[10]. D'après Hirst (1987 : 5-6), cette possibilité de caractérisation conjointe d'un mot par homonymie et polysémie démontre le flou des frontières entre les catégories proposées dans la littérature.

Un certain nombre de critères et de tests servant à distinguer la polysémie et l'homonymie, sur la base de la présence ou non de liens entre sens, ont été proposés[11]. Cependant, le phénomène de la polysémie englobe à lui seul une multitude de cas où les mots présentent des sens liés entre eux de manières variées. En faisant abstraction des cas d'homonymie, la simple distinction entre unités monosémiques et polysémiques (basée sur l'existence d'un ou de plusieurs sens) paraîtrait en effet très grossière.

Entre la modulation contextuelle (dans le cas de la monosémie) et le repérage de sens différents (dans le cas de la polysémie), il existe de nombreux degrés de distinctivité intermédiaires, dont ni la définition ni l'utilisation pour la caractérisation des mots ne semblent évidentes. Tel est le cas, par exemple, des mots à **polysémie logique** (Pustejovsky, 1995 : 90), caractérisés par Cruse (1996, 2004 : 112-115) et Kleiber (1999 : 87-94) comme des mots à **facettes sémantiques**. Cette polysémie diffère de la polysémie standard, dans la mesure où les distinctions ne concernent pas des sens différents mais des **aspects** différents de sens. Ainsi, un élément lexical peut comprendre différents sens (ou aspects de sens), de telle sorte qu'il puisse présenter à la fois la totalité des sens agglomérés

prononcés différemment, d'**homophones** (dans le langage parlé), qui sont prononcés de la même manière mais s'écrivent de manière différente, et d'**hétéronymes**, homographes non homophones.

[10] Par exemple, le mot anglais *bank* est souvent utilisé pour illustrer le phénomène de l'homonymie, en raison de ses deux sens bien distincts : « rive » et « institution financière ». Néanmoins, ce même mot peut aussi bien être caractérisé comme polysémique, puisque d'autres sens peuvent être repérés à l'intérieur même de l'un de ses sens principaux (« institution financière »), à savoir les sens « institution » et « bâtiment », entre lesquels existe un lien sémantique (Pustejovsky, 1995 : 27). L'exemple donné par Hirst (1987) pour illustrer le cas où un mot est conjointement affecté par l'ambiguïté catégorielle et la polysémie concerne le mot anglais *respect*, dont les sens nominal et verbal sont liés. Le mot *sink* est en même temps caractérisé, quant à lui, par l'ambiguïté catégorielle et l'homonymie, dans la mesure où son sens nominal et son sens verbal ne sont pas liés. Enfin, le mot *plant*, qui nous servira comme exemple dans la suite de cet exposé, est affecté par les trois types d'ambiguïté lexicale.

[11] Nous les décrirons dans le paragraphe 1.2.

et chacun d'eux[12]. Ce type de variation sémantique se situe entre la polysémie et la variation contextuelle simple. Les mots de ce type ont un contenu sémantique unitaire ou global mais présentent des composantes qui peuvent apparaître en emploi et donnent ainsi lieu à une variation de sens non polysémique et non simplement contextuelle de l'élément lexical.

La notion de **vague**, souvent liée au caractère générique des mots, occupe aussi une place importante dans la discussion sur l'ambiguïté. Selon Ullmann, les mots ne dénotent pas des éléments uniques mais des classes d'objets ou d'événements liés par un élément commun[13] (1962 : 116). Les membres d'une classe peuvent avoir des **traits non-distinctifs**, c'est-à-dire des traits qui les différencient entre eux sans pour autant remettre en cause leur appartenance à la classe, et des **traits distinctifs** qui les différencient des objets d'une autre classe. Cette conception du vague s'apparente à la notion d'**air de famille**, qui désigne les **similarités** liant des objets représentés par le même concept mais qui peuvent, en même temps, être caractérisés par des **dissimilarités** quant à différents aspects (Wittgenstein, 1953 : 27-28).

Parmi les types intermédiaires d'ambiguïté, nous trouvons aussi l'**ambiguïté sémantique faible**, qui affecte, d'après Simpson (1989), presque tous les mots du lexique. Même un mot dénotant un objet unique peut en effet être interprété de différentes manières dans des contextes différents, comme c'est le cas pour un mot ambigu, en raison de l'activation d'ensembles de ses traits pouvant être plus ou moins en contraste. Simpson considère, par conséquent, que les mots possédant un seul sens dans des dictionnaires peuvent poser le même problème que les mots « clairement » ambigus, à savoir sélectionner l'interprétation la plus adéquate du mot en fonction du contexte.

Ce raffinement des types de polysémie met en évidence de nouvelles notions qui servent à caractériser les mots par rapport à leur sémantique. Ainsi, la possibilité de collaboration entre deux acceptions d'un mot, dans le cas des facettes sémantiques, montre l'absence d'**antagonisme** entre les acceptions en

[12] Ces aspects différents des mots peuvent, selon Pustejovsky, être désignés par la notion de « Paradigme Conceptuel Lexical » (*Lexical Conceptual Paradigm*) (1995 : 91).

[13] Les exceptions à cette généralité concernent, outre les noms propres, certains noms communs qui font référence à des objets uniques ainsi que les termes scientifiques et techniques rigoureusement définis et donc privés de ce caractère générique.

question[14]. Cette notion d'antagonisme ou de concurrence est étroitement liée à la notion de **distinctivité** des sens des mots ambigus, que nous étudions ici.

La distinctivité des significations associées à une forme ambiguë décrit le degré par rapport auquel les significations se présentent comme distinctes et mutuellement exclusives. C'est pourquoi elle s'applique davantage à l'homonymie qu'à la polysémie (Fuchs, 1996 : 10-11). En cas d'homonymie, il est clair qu'à la forme ambiguë correspondent des significations totalement disjointes, puisque l'on a affaire à deux expressions constituées indépendamment l'une de l'autre, et n'entretenant entre elles aucun rapport sémantique (par ex. dans le cas du mot *avocat*). En revanche, dans le cas de polysémie, les différentes significations de l'expression sont apparentées et partagent un certain sémantisme commun[15].

Cette notion de distinctivité des acceptions d'un mot est fortement liée au degré d'antagonisme entre celles-ci, c'est-à-dire au degré de leur exclusion mutuelle[16] (Cruse, 2004 : 104-106). Les **sens** des unités lexicales polysémiques peuvent donc être **plus ou moins distincts** et **plus ou moins concurrents**. D'après Cruse, le plus haut degré d'antagonisme se manifeste entre les acceptions des mots indéniablement ambigus et représente le plus haut degré de distinctivité. L'estimation de l'antagonisme entre les sens d'un mot est donc liée, elle aussi, à l'estimation de l'existence de lien sémantique entre eux.

D'autres types de variation sémantique viennent s'ajouter aux types précités, tels ceux des **perspectives**, des **micro-sens**, des **sens locaux** et de la **généralité d'usage** (ou d'**application**). Les perspectives concernent les cas où les acceptions du mot peuvent être distinguées tout en étant unifiées dans l'esprit au sein d'une seule unité conceptuelle[17] (Cruse, *ibid*.). Les micro-sens sont des

[14] Cruse donne comme exemple, pour illustrer le phénomène des facettes sémantiques, les acceptions « texte » (contenu) et « tome » (objet physique) du mot *book* (1996 : 112-113).

[15] Cette parenté ne les empêche pas d'être distinctes du point de vue de la dénotation, dans la mesure où elles renvoient à des référents différents.

[16] Les jugements de distinctivité entre les acceptions d'un mot peuvent être fondés sur un ensemble de critères comme l'identité de référence, l'indépendance des conditions de vérité, l'indépendance des ensembles de relations sémantiques des acceptions (par ex., relations de synonymie ou d'antonymie) ainsi que leur autonomie, c'est-à-dire la possibilité d'utilisation du mot dans un de ses sens quand l'autre est explicitement éliminé.

[17] Ainsi, le sens de *book* dans « John began the book », qui a deux acceptions possibles : « John a commencé à lire un livre » ou « il a commencé à l'écrire ». Dans les deux cas, la facette TEXTE étant impliquée, l'ambiguïté ne peut être expliquée par recours aux facettes. Ce type de variation

acceptions caractérisées par un degré plus bas de distinctivité et d'antagonisme que les sens. Ils fonctionnent en tant que sens au sein de leur domaine spécifique, mais sont moins accessibles hors du domaine[18]. Les sens locaux, s'ils relèvent également d'un domaine, diffèrent des micro-sens sur certains aspects : il s'agit de points situés sur un continuum sémantique (**sense spectrum** dans Cruse, 1986) ; leur degré d'antagonisme dépend de la distance qui les sépare dans le spectrum ; l'induction de sens littéraux et figuratifs est possible mais non une acception inclusive[19]. Enfin, la généralité d'usage – autrement dit, la microdistinction – concerne les termes considérés comme « vagues ».

Ce raffinement très poussé des types de variation sémantique, s'il rend possible la description de nombreux phénomènes, augmente la difficulté à distinguer les types et à caractériser les mots par rapport aux types. Ainsi, distinguer les cas d'ambiguïté de ceux de microdistinction n'est pas toujours aisé, car les principes sous-jacents sont souvent flous et mal définis[20]. Des expériences menées en psycho-linguistique (Jorgensen, 1990) ont démontré cette difficulté à appréhender les distinctions : les mêmes cas étant caractérisés, par certains sujets, comme des cas d'ambiguïté et, par d'autres, comme des cas de microdistinction. Geeraerts (1993) souligne également la difficulté à distinguer entre la polysémie et le vague. La validité des critères utilisés pour établir une telle distinction se trouve remise en cause du fait qu'ils peuvent, d'une part, entrer en conflit et donc empêcher d'aboutir à la même conclusion dans les mêmes circonstances et que chacun de ces critères peut, d'autre part, générer des résultats différents relativement à la distinctivité d'un sens lexical dans deux contextes différents.

sémantique concerne un certain niveau de distinctivité sans antagonisme, mais à un degré différent des mots à facettes et sans autonomie.

[18] Tels sont, par exemple, les différents types de couteau, qui sont tous des couteaux, et possèdent une acception par défaut ainsi que des acceptions plus spécifiques (couteau pour manger, outil, arme, instrument chirurgical).

[19] Un exemple donné par Cruse, pour cette catégorie de mots, concerne le mot *mouth* qui a un sens basique : la « bouche » d'un humain ou d'un animal. Les autres acceptions en sont des extensions métaphoriques (comme "mouth of river").

[20] Se trouve, chez Jorgensen (1990), une critique de la tentative de Katz et Fodor (1963) pour formaliser un modèle de sens lexicaux expliquant le processus de désambiguïsation. Ce processus utilise un métalangage de restrictions de sélection nécessaires et suffisantes pour décrire chaque sens qu'un mot peut prendre en contexte. Or l'absence d'un principe définissant les différences dans le contexte d'utilisation qui provoquent des différences sémantiques importantes dans le sens d'un mot, explique qu'une telle tentative implique la description d'un nombre illimité de sens (autant qu'il y a de contextes d'occurrence ou de valeurs de vérité) ; de telle manière que le nombre de restrictions de sélection nécessaires est égal au nombre de distinctions qu'elles servent à décrire.

La catégorisation des mots par rapport aux différents types d'ambiguïté n'est donc pas, à l'évidence, une tâche simple. D'une manière générale, analyser et caractériser la sémantique des mots ambigus implique des considérations quant au repérage des relations éventuellement entretenues entre leurs sens, ainsi qu'à l'estimation de la distinctivité et de l'antagonisme qui peuvent exister entre eux. Les différents cas d'ambiguïté pourraient, par conséquent, être illustrés à l'aide d'un **continuum** allant des cas d'indétermination sémantique, où une distinction entre les sens (ou les nuances de sens) du mot n'est pas évidente, à ceux qui impliquent l'existence de sens à ce point distincts qu'aucune relation ne peut être repérée entre eux (tels que les cas d'homonymie). A l'intérieur de ce gradient, pourraient se situer les unités lexicales qui présentent des distinctions sémantiques plus ou moins claires et de granularité variée[21], caractérisées par un degré plus ou moins grand d'ambiguïté[22].

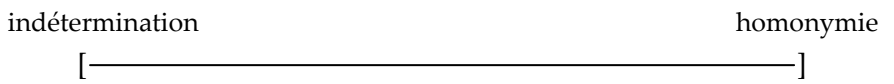


Figure 1. Intervalle fermé d'ambiguïté

Il ne faut néanmoins pas exclure la possibilité de déplacement d'une unité lexicale au sein du continuum d'ambiguïté. Comme le souligne Ullmann (1962 : 160), dans le cas du vague, les mots ont un nombre d'aspects différents selon les contextes dans lesquels ils sont utilisés. Certains de ces aspects sont éphémères, tandis que d'autres peuvent se développer en nuances de sens permanentes et, si l'écart entre eux s'élargit, ils peuvent même être considérés comme des sens différents du même terme[23]. Bien que ces différentes étapes soient distinguées de

[21] Ce type de variation est bien décrit par les **qualia roles** proposés par Pustejovsky (1995).

[22] La notion de **continuum** peut aussi être appliquée à la sémantique d'une unité lexicale. Les sens véhiculés par l'unité peuvent être décrits à l'aide d'un espace sémantique continu (sans frontières entre eux), dont une région est activée lors de la mise en contexte de l'unité. Une telle représentation ne rend pas nécessaire la délimitation stricte des sens. La taille de la région attribuée à une unité permet la description de sens indéterminés (si la région est étendue) ou plus précis (si la région est étroite), ainsi que la description de l'ambiguïté (si la région est constituée de parties disjointes). L'activation d'une région de l'espace, lors de la désambiguïsation d'une unité, permet aussi le partage de la même région par des occurrences de l'unité dans des énoncés différents (Victorri et Fuchs, 1996 : 67-70).

[23] « *As we saw when we discussed the various forms of vagueness in meaning, our words have a number of different aspects according to the contexts in which they are used. Some of these aspects are purely ephemeral;*

manière systématique dans les dictionnaires, dans la réalité, des fusionnements imperceptibles ont lieu entre eux. Ainsi la polysémie des mots peut évoluer et des mots vagues peuvent progressivement devenir polysémiques. Il arrive même que des mots restent stables par rapport à certains de leurs aspects alors que des distinctions claires s'établissent dans une autre région de leur espace sémantique.

Le même continuum pourrait être conçu comme un intervalle semi-ouvert (ouvert à gauche), dont l'extrémité gauche désigne les cas de **monosémie**, autrement dit des cas d'**univocité sémantique**.

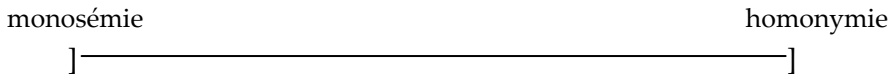


Figure 2. Intervalle semi-ouvert d'ambiguïté

Les cas de monosémie sont rares dans les langues naturelles et ceux qui existent apparaissent dans des vocabulaires spécialisés (terminologies), employés dans des domaines bien précis. D'après la définition de l'ISO (Lerat, 1995), la monosémie est la « relation entre désignation et notion dans laquelle une désignation représente une seule notion »[24]. La monosémie est effectivement une notion omniprésente dans les travaux portant sur les langues de spécialité et fonctionne d'ailleurs généralement comme pierre de touche pour différencier ces langues de la langue générale. Elle a été même considérée pendant longtemps comme une « vertu » dont les langues générales seraient privées (Lerat, *ibid*. : 90-91). Cette dichotomie entre langue générale et langue spécialisée est pourtant rejetée par les partisans de la terminologie descriptive (Condamines et Rebeyrolles, 1997). En outre, des travaux récents (Bertels, 2005 ; Bertels *et al*., 2006) démontrent l'absence de corrélation entre **continuum de spécificité** et **continuum de polysémie**, signifiant ainsi qu'un fort degré de spécialisation des mots ne sous-entend pas un degré aussi fort de monosémie. Cette conclusion va à l'encontre du point de vue de la terminologie traditionnelle sur la sémantique

others may develop into permanent shades of meaning and, as the gap between them widens, we may eventually come to regard them as different senses of the same term. In the dictionaries, these various stages are systematically distinguished, but in actual fact they merge imperceptibly into one another. » (Ullmann, *ibid*.: 160).

[24] Voir aussi ISO/IEC Directives, Supplement-Procedures specific to IEC, Third Edition, 2008, p. 26.

des unités des langues spécialisées, qui défend leur univocité, leur monoréférentialité et leur monosémie. Elle remet même en cause la nécessité de distinguer mots de langue générale et mots spécialisés, signalant par là que les mots ont probablement tous besoin du même type de traitement.

Après avoir présenté un certain nombre de problèmes liés à la catégorisation des mots par référence à leur sémantique, nous allons désormais analyser un facteur qui joue un rôle important dans la caractérisation de la sémantique des mots. Il s'agit des relations entretenues entre les sens, dont le repérage est crucial pour pouvoir caractériser les mots relativement aux différents types d'ambiguïté.

1.2. Analyse des relations entre les sens

La caractérisation des mots par rapport aux types d'ambiguïté proposés pose souvent le problème du repérage des relations éventuellement existantes entre les sens des mots et celui de l'identification de la nature des relations entretenues. Dans une **perspective historique**, l'existence de relations de ce type peut être étudiée en faisant appel à l'étymologie des mots. D'un **point de vue diachronique**, les homonymes ont des étymologies distinctes, tandis que les différentes significations d'un polysème correspondent toutes à la même unité d'origine (*étymon*) (Martinet, 1974 ; Lyons, 1990 ; Fuchs, 1996 : 26-27). Ce type de critère, basé sur l'étymologie des mots, arrive habituellement à éclairer la relation entre leurs différents sens. Néanmoins, dans une perspective synchronique, la référence à l'histoire est exclue (Martinet, 1974 ; Polguère, 2003 : 51). Dans ce cas, il faut d'autres critères pour repérer la présence ou l'absence d'un lien sémantique entre les sens ainsi que pour caractériser les mots en fonction de leur ambiguïté.

Dans le cadre d'une **approche synchronique**, un lien étymologique qui ne se concrétise plus par une relation de sens couramment perçue par les locuteurs doit être ignoré (Polguère, *ibid*.). Il arrive ainsi que des mots ayant la même étymologie soient considérés comme distincts quand, du point de vue d'un locuteur ordinaire, il n'y a pas d'analogie claire entre leurs usages (Quine, 1960 : 129). Ce type de critère synchronique est caractérisé par Fuchs (1996 : 26-27) comme un **critère d'ordre épi-linguistique**, qui a trait à la façon non théorisée

dont les locuteurs réagissent spontanément à la langue. Ainsi, dans une approche épi-linguistique, les formes traitées comme des homonymes sont celles dont les sujets parlants s'accordent spontanément à considérer les significations correspondantes comme étrangères les unes aux autres, tandis que les formes considérés comme polysémiques sont celles à propos desquelles les sujets perçoivent une certaine parenté entre les différentes significations.

A ce type de critères synchroniques s'ajoute celui des **critères synchroniques d'ordre théorique** (Fuchs, *ibid.*)[25]. D'un point de vue théorique en synchronie, on parle d'unités homonymes lorsque les représentations sémantiques construites pour rendre compte des différentes significations sont totalement disjointes. On parle, en revanche, d'un polysème, lorsque l'on est capable de représenter la parenté des significations, en termes d'éléments de sens communs, de dérivation ou d'analogie de sens. La distinction entre homonymes et polysèmes peut, par exemple, être établie à l'aide de la notion de sèmes, en ce sens que la polysémie suppose une cohérence sémique qui manque aux homonymes (Picoche, 1984 ; Lyons, 1990). Les liens de sens unissant les divers signifiés d'une même forme polysémique peuvent aussi être définis en utilisant des principes sémantico-logiques de dérivation (Condamines et Rebeyrolles, 1997). Ainsi, la polysémie lexicale peut être décrite en considérant les modifications qui peuvent s'opérer sur le sème central d'un mot, par des glissements de sens du type métaphore, métonymie, spécialisation, extension, restriction, etc.

Un ensemble de tests appliquant les critères théoriques en synchronie a été proposé afin de rendre possible l'analyse des liens sémantiques et des différences entre les sens des mots ambigus. Par exemple, le **test de substituabilité** (Victorri, 1997) rend possible l'estimation de la différence entre les sens d'un mot polysémique par sa substitution avec d'autres mots dans des expressions différentes. La différence de sens est mise en évidence par le fait que le mot étudié peut être remplacé par certains mots dans certaines expressions, sans provoquer de changement important de sens, alors que pour d'autres expressions, il faut faire appel à d'autres termes afin que le sens initial subsiste.

[25] La convergence des critères d'ordre diachronique et des critères épi-linguistiques et théoriques en synchronie facilite la prise de décision. Pourtant, des divergences peuvent apparaître, surtout entre critères synchroniques et critères d'ordre diachronique (Fuchs, *ibid.*).

La mise en évidence de mots par lesquels le mot polysémique est remplaçable dans toutes les expressions ou d'une suite de sens ressentis comme « intermédiaires » dans certaines expressions peuvent servir à dégager la parenté ou l' « air de famille » (Wittgenstein, 1953) qui lie deux sens.

L'estimation théorique de la parenté entre les sens concerne également des **critères morphologiques** (Condamines et Rebeyrolles, 1997). Ainsi, si les différents sens correspondent à des constructions différentes et donnent lieu à des dérivés spécifiques, ils ne sont pas liés et peuvent alors être attribués à des termes homonymes. Dans le cas inverse, ils sont attribués à une unité polysémique.

Des méthodes permettant l'identification des relations entre les sens des mots polysémiques ont été également proposées dans un **cadre automatique**. Ces méthodes concernent surtout l'identification de liens entre les sens répertoriés au sein de dictionnaires et d'autres inventaires sémantiques (Chodorow, 1990 ; Dolan, 1994 ; Peters *et al*., 1998 ; Gonzalo *et al*., 2000 ; Mihalcea et Moldovan, 2001 ; Pustejovsky, 1995). La méthode proposée par Chodorow, par exemple, concerne l'identification automatique d'instances de certains types récurrents de relations sémantiques (processus/résultat, nourriture/plante, etc.), caractérisés comme des cas de polysémie logique. L'identification automatique des relations de ce type permettrait la construction automatique de bases de données lexicales, où la manière dont les différents sens des mots sont liés serait explicitement montrée[26]. Les autres travaux cités ci-dessus s'intéressent au repérage de liens sémantiques au sein de ressources sémantiques décrivant des sens de granularité très fine (comme WordNet), qui rendrait possible le regroupement de ces sens et la diminution de la granularité des descriptions. Nous reviendrons sur ces travaux dans le paragraphe consacré aux avantages et aux inconvénients de l'utilisation de ressources préexistantes pour la désambiguïsation lexicale dans un cadre automatique (§2.3, chapitre 3).

[26] Chodorow souligne que l'identification automatique d'instances de polysémie systématique au sein des dictionnaires électroniques ne permet de capter qu'un petit sous-ensemble de cas, dans lesquels les sens des mots se recouvrent. La raison en est le grand nombre de distinctions fines et idiosyncrasiques des mots, qui ne peuvent être caractérisées d'une manière générale. La prise en compte de l'ensemble des sens lexicaux liés nécessiterait un codage manuel fastidieux, ce qui explique, selon Chodorow, l'absence de tentatives de grande envergure visant la création automatique de liens étiquetés entre les sens.

La variation des conceptions du sens constitue une source de divergences entre ressources différentes au niveau des descriptions sémantiques. De notre côté, nous avons souligné le rôle joué par le repérage des relations inter-sens et par l'analyse de la nature de ces relations sur la caractérisation des mots quant à leur ambiguïté. Pourtant, ces questions sont généralement négligées lors de la création d'inventaires sémantiques. Comme nous le montrerons de façon plus approfondie dans le chapitre consacré à la présentation de la méthode d'analyse sémantique proposée au sein de ce travail (chapitre 6), cette méthode met en évidence de manière automatique les relations entretenues entre les sens. Nous présenterons également, dans le neuvième chapitre de cette thèse, une méthode d'identification automatique de la nature de ces relations.

Dans le paragraphe suivant, nous allons présenter les facteurs théoriques et extra-linguistiques qui interviennent lors de la création de ressources sémantiques. Ces facteurs influencent en effet l'énumération des sens au sein d'inventaires sémantiques et contribuent, par conséquent, au manque d'uniformité observé entre ressources différentes.

1.3. Enumération des sens au sein d'inventaires sémantiques

1.3.1. Influence des conceptions théoriques du sens

Le besoin d'**énumération des sens** et de définition de **critères** clairs qui assureraient une bonne répartition entre polysèmes et homographes est manifeste dans la pratique lexicographique, dans la mesure où il s'agit précisément de créer des inventaires lexicaux où les sens doivent être répertoriés et bien décrits. La pratique lexicographique moderne implique l'utilisation de **corpus de textes réels** : le lexicographe observe les occurrences des mots trouvées dans les textes, les regroupe sur la base de leur similarité, définit leurs éléments communs et encode ses conclusions en guise de définitions. Ce regroupement d'instances de mots peut être considéré comme un cas de **clustering** basée sur les données (Pedersen, 2007).

Les critères sous-jacents à cette clusterisation ne sont pas toujours clairs et il se peut que, même le lexicographe, n'en soit pas conscient (Kilgarriff, 1997c,

2006). Par conséquent, les distinctions sémantiques établies dans les dictionnaires sont souvent arbitraires (Atkins, 1991 ; Atkins et Levin, 1998) ; ce qui constitue l'une des raisons expliquant les **divergences** constatées entre ressources différentes. Ces divergences se manifestent souvent aux niveaux du nombre des entrées dictionnairiques et de la répartition des sens en leur sein. Ainsi, des sens différents d'un mot sont tantôt décrits par une seule entrée dictionnairique, tantôt distingués et décrits par des entrées différentes – phénomène appelé **multiplication des homonymes**. Les raisons sous-jacentes à ces divergences peuvent être de diverse nature. Les raisons sémantiques qui interviennent sont liées aux questions discutées dans le paragraphe précédent et qui concernent la possibilité de délimitation des sens. Nous verrons, par la suite, quelle est leur influence au niveau de la pratique lexicographique.

Le phénomène de la multiplication des homonymes est observé lorsque la parenté entre les significations des expressions polysémiques n'est pas prise en compte. Fuchs parle, dans ce cas, de **polysémie éclatée** : « *la polysémie d'une expression est éclatée en homonymie, en ramenant les cas d'indétermination du sens à des cas d'ambiguïté et forçant, en termes d'alternatives, toutes les incertitudes interprétatives* »[27] (1996 : 32-34). A l'opposé de cette approche, nous trouvons l'approche de la **polysémie réduite**[28] : « *par souci d'éliminer l'idée de polysémie, les significations correspondantes à une forme polysémique sont considérées comme sur-déterminées, évacuées hors du champ de la langue, et souvent traitées comme de simples « effets de sens en discours »* ». Un noyau de sens unique en langue et largement sous-déterminé est substitué à ces significations, noyau censé être sous-jacent à la diversité des significations en contexte (Fuchs, *ibid*.).

Pour la polysémie éclatée, toute indétermination se radicalise en ambiguïté, tandis que pour la polysémie réduite, toute ambiguïté se dissout en indétermination. Les mêmes données linguistiques peuvent donc être décrites en termes d'univocité (en sous-déterminant le sens de l'expression en question) ou en termes d'ambiguïté (en sur-déterminant le sens de l'expression). La variabilité

[27] Attitude souvent adoptée par les écoles de linguistique formelle et informatique, qui pratiquent l'éclatement homonymique systématique et sur-déterminent la signification en contexte.

[28] Approche ayant ses racines dans le structuralisme, qui défend la bi-univocité des rapports entre formes et sens ; elle consiste à réduire la polysémie à une sorte d'univocité sous-déterminée, ramenant ainsi les cas d'ambiguïté à des cas d'indétermination.

des approches possibles face à l'ambiguïté souligne la difficulté d'atteindre un consensus sur le sujet[29]. L'absence de critères clairs justifiant ces choix ramène le traitement du problème à un autre niveau, celui de l'attitude adoptée par les lexicographes. Ainsi, Wilks et Stevenson (1996) répartissent les lexicographes en deux catégories : les **diviseurs** (*splitters*, en anglais), qui choisissent de diviser les sens sans fin et les **fusionneurs** (*lumpers*, en anglais), qui préfèrent des clusters d'usages plus larges ou, autrement dit, plus « homographiques ». Pour Kilgarriff (1997c), tout lexicographe professionnel doit prendre quotidiennement des décisions inévitablement subjectives à propos de la « division » des sens (séparation de deux motifs d'usage légèrement différents en des sens différents) ou de leur « fusion » (considération de deux motifs d'usage légèrement différents comme un seul sens).

D'un point de vue sémantique, les divergences observées entre ressources à ces niveaux peuvent être liées aux variations de la conception de la nature du sens, des types de polysémie et des relations entretenues entre les sens lexicaux. Ces variations de conception sont liées à la question de l'**unicité** de la forme ambiguë, question liée, à son tour, à la distinctivité des significations associées à la forme ambiguë et à leur antagonisme[30]. En ce qui concerne l'unicité, dans le cas de l'**homonymie**[31]**,** l'ambiguïté est due à « *la collision accidentelle entre les formes de deux signes linguistiques distincts. (…) Dans ces cas, le dédoublement en deux unités – dont témoigne (…) l'existence de deux entrées lexicales distinctes dans les dictionnaires – permet de retrouver une correspondance biunivoque entre formes et sens* » (Fuchs, 1996 : 9). Suite à ce **dédoublement**, chaque entrée est traitée comme univoque, c'est-à-dire comme ne possédant qu'une seule signification. En revanche, dans le cas de la **polysémie**, on ne parle que d'**une seule unité** polysémique, caractérisée par un rapport de non-biunivocité entre forme et sens[32].

Les divergences quant à la nature des sens des mots polysémiques et les relations qu'ils entretiennent, qui sont relatives aux aspects de l'unicité, de la

[29] Certains cas d'ambiguïté sont généralement admis, comme les ambiguïtés dues à la présence de formes homonymes ou celles qui mettent en jeu des significations référentiellement disjointes d'expressions polysémiques.

[30] Les notions de distinctivité et d'antagonisme ont été analysées dans le paragraphe 1.1.

[31] L'homonymie regroupe, pour Fuchs, les cas d'homophonie et d'homographie.

[32] Dans le cas de la polysémie, « on a affaire à une seule et même expression, à un unique signe linguistique ; c'est alors l'expression elle-même qui est ambiguë, dans la mesure où elle possède une forme à laquelle correspond une pluralité de significations. » (Fuchs, *ibid*.).

distinctivité et de l'antagonisme, sont habituellement négligées dans la pratique lexicographique et ne sont pas reflétées au sein des descriptions sémantiques fournies par les dictionnaires. L'absence d'informations sur le lien entre les différents sens des mots est d'ailleurs souvent considérée comme une faiblesse des dictionnaires existants (Dolan, 1994[33]). Elle est aussi parfois attribuée aux techniques de représentation et aux stratégies de différenciation des sens utilisées, incapables de justifier une distinction entre les différents types d'ambiguïté (Pustejovsky, 1995 : 29). Par exemple, les **Lexiques d'Enumération de Sens**, qui constituent une approche largement répandue de description des sens des mots, ne permettent pas la prise en compte de la relation logique existante entre eux (comme dans le cas de la polysémie complémentaire, *ibid.* : 37), ni la description d'aspects comme l'usage créatif des mots, la perméabilité des sens lexicaux et la possibilité de réalisations syntaxiques multiples d'un sens lexical. Ce type de représentation[34] permet seulement, d'après Pustejovsky, la description des cas d'ambiguïté restreinte par la pragmatique (c'est-à-dire, les cas d'ambiguïté contrastive). L'adoption d'une telle approche pour la description des cas de polysémie complémentaire entraînerait une prolifération des sens lexicaux.

1.3.2. Influence des facteurs extra-linguistiques

Nous avons déjà souligné qu'outre les **facteurs sémantiques** affectant l'organisation des entrées dictionnairiques, d'autres types de facteurs peuvent aussi intervenir et conditionner les choix effectués lors de la création d'un dictionnaire. Gale *et al*. (1993) présentent une liste de facteurs non sémantiques qui influencent la structure d'un dictionnaire et le découpage des entrées. Ces facteurs concernent les différences au niveau des parties de discours, des traits syntaxiques (comme comptable/massif, personne, nombre, genre, etc.), des structures de valence (verbes transitifs/intransitifs), de la prononciation (rarement et sans que cela constitue l'unique raison régissant la répartition des

[33] Dolan distingue quatre types de relations possibles entre les sens : les sens qui diffèrent par rapport à des nuances sémantiques très fines, ceux qui sont liés historiquement mais non du point de vue synchronique, ceux qui sont liés par un processus, plus ou moins opaque, de métaphore ou de métonymie, et enfin ceux qui semblent n'être absolument pas liés.

[34] A l'aide de listes de mots dont chacun est annoté par un sens distinct.

sens), de l'étymologie (rarement dans des dictionnaires pour apprenants, plus fréquemment dans des dictionnaires basés sur des principes historiques), de la capitalisation, du registre, du dialecte, des collocations et des informations de domaine (souvent dans les versions électroniques).

A cette liste de facteurs **non-sémantiques** s'ajoutent ceux qui pourraient être caractérisés comme **extra-linguistiques**. De tels facteurs sont, par exemple, la finalité du dictionnaire et le public visé[35]. L'influence des facteurs extra-linguistiques sur l'organisation des dictionnaires a déjà été soulignée par Quine (1960 : 129), selon lequel lexicographes et grammairiens établissent des distinctions de mots à leur convenance, au-delà des injonctions de la forme et de l'étymologie. Ainsi, lorsque le dictionnaire est conçu dans un but de traduction, la division en homonymes peut répondre au besoin de disposer de deux mots distincts dans la LC, qui couvrent l'espace couvert par un seul mot source[36]. Cette particularité structurale des dictionnaires bilingues est également soulignée par Brun *et al*. (2001), selon qui les impératifs de correspondance avec l'autre langue rendent le découpage en sens dans chaque langue plus systématique[37]. Les dictionnaires de ce type sont donc caractérisés par une granularité sémantique plus fine que les dictionnaires monolingues[38].

La prise en compte du **public visé** a, par ailleurs, un impact important au niveau de la structuration des dictionnaires. Un dictionnaire pour enfants ne contiendra pas la même quantité d'informations qu'un dictionnaire pour adultes et même la nature des informations incluses y sera différente. Par conséquent, un dictionnaire bilingue destiné à être utilisé par des enfants qui apprennent une langue étrangère ne contiendra pas le même type d'informations qu'un

[35] Kilgarriff (1997c) inclut aussi les stratégies éditoriales comme l'un des facteurs qui influencent la construction des dictionnaires et les distinctions sémantiques qui y sont décrites.

[36] Pour Quine, la multiplication des homonymes a beau rendre les choses parfois plus faciles, elle crée néanmoins des problèmes supplémentaires au niveau de l'identité lexicale et du concept du mot.

[37] En ce qui concerne la structure des dictionnaires bilingues, Brun *et al*. (*ibid*.) remarquent en outre que les définitions des sens sont souvent accompagnées d'un nombre d'exemples significatifs plus important que dans les dictionnaires monolingues.

[38] Les critiques de la granularité fine au sein des ressources monolingues sont très fréquentes dans la littérature. Nous estimons que la remarque faite par Brun *et al*. (*ibid*.) ne concerne que les distinctions sémantiques dont le repérage facilite l'établissement de correspondances inter-langues et qui ne seraient pas nécessaires dans une description monolingue (non incluses, par conséquent, dans les ressources correspondantes), mais nous doutons de la validité générale de cette remarque.

dictionnaire monolingue, ou qu'un dictionnaire bilingue destiné à être utilisé par des traducteurs professionnels.

Si le paramètre de la **finalité** a une influence importante sur la structure et le contenu des inventaires lexicaux destinés à être utilisés par des humains, cette importance s'accroît lorsqu'il s'agit de la conception et de la création de ressources destinées à être utilisées dans des applications automatiques. Dans un **cadre automatique**, l'approche adoptée pour l'analyse sémantique, les critères sur lesquels se base la répartition des sens ainsi que leur description dépendent fortement de l'application visée. Cependant, les besoins des applications en matière de désambiguïsation diffèrent sensiblement et doivent donc être pris en considération lors du développement des ressources, et ce, afin d'assurer la performance optimale des processus de désambiguïsation. A cette exigence d'**adéquation** des ressources sémantiques aux besoins des applications particulières s'ajoute une autre contrainte, liée au caractère automatique des tâches qui doivent être accomplies : étant donné que les informations sémantiques sont destinées à être traitées par la machine, elles doivent être **claires** et **bien définies**. La raison sous-jacente à cette contrainte réside dans le fait que les possibilités de généralisation et d'inférence dans un cadre automatique sont beaucoup plus limitées que lors du traitement des informations sémantiques par des humains. Les besoins de clarté et de précision conditionnent ainsi de manière importante la structure des ressources développées pour être utilisées dans un cadre automatique et expliquent, jusqu'à un certain point, l'inadéquation des informations incluses dans des ressources destinées à des utilisateurs humains aux applications du TAL. Cette inadéquation, constatée à plusieurs reprises, qui va de la difficulté jusqu'à l'impossibilité d'exploiter ces informations, est ainsi imputable tant au facteur de finalité qu'à celui des différences entre type de traitement.

Même en se limitant au TAL, il n'est pas possible de concevoir une ressource qui soit valable pour différentes applications. Cette impossibilité s'explique par le grand nombre et la nature variée des facteurs qui interviennent lors de la conception et de la réalisation des applications. D'après Kilgarriff (1997c), la différence des besoins d'analyse sémantique selon les applications rend

impossible la conception d'un ensemble unique de sens, indépendant d'une tâche précise et adéquat à des applications et à des tâches différentes.

L'étape de la **désambiguïsation lexicale** constitue une **étape intermédiaire de traitement** au sein des applications et non un but en soi. L'objectif visé par l'application détermine donc fortement les exigences en matière de désambiguïsation. Ainsi, les besoins en matière de désambiguïsation d'une application opératoire dans un cadre bilingue ou multilingue peuvent être satisfaits par un inventaire sémantique contenant seulement les distinctions lexicalisées dans l'(les) autre(s) langue(s) (Resnik et Yarowsky, 1997). Plus concrètement, dans une application de traduction, où l'objectif de la désambiguïsation est la prédiction de la traduction correcte de nouvelles occurrences des mots polysémiques de la LS, la distinction entre les sens véhiculés par les mots source ambigus n'est pas toujours nécessaire, en raison du phénomène de préservation de l'ambiguïté (ou d'une partie de celle-ci) dans l'autre langue[39] (Edmonds et Kilgarriff, 2002). Une ressource sémantique décrivant uniquement les cas où les sens des mots source sont lexicalisés de manière différente dans la LC peut suffire dans le cadre d'une telle application, dans la mesure où des prédictions de traduction correctes peuvent être faites sans nécessairement passer par l'étape de désambiguïsation[40]. En revanche, dans le cas d'applications comme la recherche d'information (monolingue et bi-(multi-)lingue) à partir du Web, les besoins de désambiguïsation diffèrent sensiblement. La nécessité de résoudre l'ambiguïté des mots polysémiques inclus dans les requêtes est cruciale pour obtenir des résultats de bonne qualité et éliminer le bruit inclus dans les résultats.

[39] Ce phénomène, appelé **ambiguïté traductionnelle** ou **ambiguïté parallèle**, sera analysé plus loin. Nous donnons ici seulement un exemple : le mot *interest* en anglais et le mot *intérêt* (qui constitue un candidat de traduction très probable en français) véhiculent tous les deux les sens « intérêt financier » et « intérêt personnel ». La traduction correcte en français d'une nouvelle occurrence de ce mot ne nécessite donc pas obligatoirement la désambiguïsation de cette nouvelle occurrence

[40] L'idée de Resnik et Yarowsky est de définir un ensemble de langues cibles et de dictionnaires bilingues associés et de requérir que chaque distinction sémantique soit lexicalisée dans un sous-ensemble minimum de ces langues. Cette stratégie éliminerait beaucoup de distinctions, mieux traitées en tant que polysémie régulière. Les informations incluses dans une telle ressource se situent à mi-chemin entre la restriction aux homographes dans une seule langue (qui donnerait des distinctions de granularité très grossière) et l'expression de toutes les distinctions de granularité fine, comme c'est le cas dans les dictionnaires monolingues.

La question de l'utilisation de ressources préétablies dans les applications du TAL sera traitée en détail dans le paragraphe 2.3. du chapitre 3., où les hypothèses sous-jacentes à une telle tentative seront analysées, ainsi que les avantages et les inconvénients qui lui sont liés.

2. La polysémie lexicale dans la traduction

2.1. Considérations théoriques autour du découpage sémantique des langues

La tentative de mise en correspondance des unités lexicales de deux langues rend souvent explicites les différences qui existent au niveau du système des langues en question. Ces variations peuvent être considérées comme des effets de la manière différente dont les langues découpent leur **espace sémantique**. Le monde conceptuel évolue, effectivement, d'une manière propre à chaque langue, pour des raisons historiques, culturelles, géographiques ou sociales, et les divergences observées ne concernent pas seulement le nombre de concepts encodés dans le vocabulaire mais aussi la structure des systèmes conceptuels. Une hypothèse théorique assez forte visant à proposer une explication à la variété des découpages sémantiques opérés par les langues naturelles est celle de la **relativité linguistique**, initialement proposée par Boas (1911), reprise par Sapir et Whorf et transformée en hypothèse du **réductionnisme culturel**[41] (Sapir, 1921 ; Whorf, 1956). D'après l'hypothèse de la relativité linguistique, chaque langue représente une classification différente de l'expérience. Pour Boas, les classifications linguistiques reflètent celles de la pensée ; ainsi, les données linguistiques peuvent constituer des matériaux utiles à l'étude des idées conceptuelles et des formes de pensée caractéristiques d'une culture. L'hypothèse du réductionnisme culturel considère, quant à elle, la langue, la culture et la pensée comme des miroirs les unes des autres ; ainsi, les sens exprimés par une

[41] Connue aussi comme l'hypothèse Sapir/Whorf. L'hypothèse de « relativité linguistique », proposée par Boas, constitue une version plus faible du « réductionnisme culturel ».

langue reflètent le contexte culturel dans lequel elle apparaît. Les distinctions linguistiques reflètent des distinctions culturelles et celles-ci génèrent des distinctions dans la pensée. Contrairement à Boas, pour lequel la langue reflète la pensée et la culture tout en n'ayant sur elles qu'une influence occasionnelle, Sapir a appréhendé la langue comme un facteur puissant de formation de la pensée, en raison de l'impact de son utilisation dans l'interprétation de l'expérience.

L'hypothèse du réductionnisme culturel a fait l'objet de nombreuses critiques. Un des problèmes, de type méthodologique, soulevé par cette hypothèse consiste en ce qu'elle présuppose la compréhension et la comparaison de langues et de cultures différentes. Or, si la compréhension se limite à une langue et à une culture, il n'est alors possible de connaître que ce qui relève d'une langue et d'une culture, la nôtre (Frawley, 1992 : 45-46). Des critiques ont également été formulées contre des exemples bien connus qui servaient de « pilier » à l'hypothèse de la relativité linguistique. Tel l'exemple à propos de la richesse du vocabulaire des Eskimos pour désigner « la neige », vocabulaire supposé être largement supérieur à celui des autres langues du monde pour parler de la même réalité[42]. Cet argument servait à démontrer la relativité des systèmes conceptuels, la diversité des catégorisations de l'expérience selon les cultures, ainsi que la division d'une sphère conceptuelle – correspondante à un seul mot dans d'autres langues – dans des classes distinctes. L'argument, repris et modifié dans des essais ultérieurs[43], peut aisément être réfuté en raison de l'absence d'une conception unique de la langue des Eskimos[44] et de la perception de ce qu'est un mot[45] (Martin, 1986 ; Pullum, 1991). Le mal fondé de cette

[42] La source de cet argument a été le texte de Boas (1911) où il parlait de l'existence de quatre racines utilisées par les Eskimos pour désigner la neige. Il a toutefois signalé que les Eskimos utilisent ces racines différentes à propos de la neige, de la même manière que les anglais utilisent des racines distinctes pour parler des formes de l'eau (*liquid, lake, river, brook, rain, dew, wave, foam*) qui, dans une autre langue, pourraient être exprimées par des dérivations morphologiques d'une racine unique signifiant « eau ».

[43] Ces essais présentent de grandes divergences quant au nombre de mots utilisés par les Eskimos, allant parfois jusqu'à 400 (Martin, *ibid.*, Pullum, *ibid.*).

[44] Les langues parlées par les Eskimos en des lieux différents (Sibérie, Alaska, Canada, Groënland) diffèrent beaucoup au niveau des détails du vocabulaire. Ainsi, il n'est pas correct de parler d'une seule langue des Eskimos, et d'émettre des jugements qui n'ont pas été vérifiés par des locuteurs des dialectes particuliers.

[45] Les langues des Eskimos sont hautement flexionnelles, avec une morphologie dérivationnelle très productive. Un exemple, cité par Pullum (1991 : 169), concerne le mot *igluksaq* considéré, dans une liste de mots d'un dialecte Eskimo propres à la neige, comme signifiant « snow for igloo making ». Ce mot est une formation productive de *iglu* (« maison ») et *–ksaq* (« matériel pour ») et, ainsi, le

hypothèse repose sur une fausse conception du vocabulaire des autres langues pour décrire la réalité « neige »[46], et peut être renforcé par des arguments sur la richesse du vocabulaire propre à certaines professions et domaines (par ex. les botanistes), qui ne démontre pas pour autant l'existence d'un lien entre la langue, la pensée et la culture.

Autre argument à l'encontre de l'hypothèse du réductionnisme culturel : la capacité des langues à transmettre le même contenu par des moyens différents. Ainsi, la non lexicalisation dans une langue du contenu d'un élément lexical existant dans une autre langue ne démontre pas l'existence de différences au niveau de la pensée et des cultures correspondantes. Un même contenu peut en effet être exprimé à l'aide d'une paraphrase, qui se situe à un niveau autre que les éléments lexicaux[47]. La question de lexicalisation des sens dans des langues différentes doit être, par conséquent, liée à la problématique du « niveau lexical ».

2.2. La question du niveau lexicale

Selon Mauranen (2002), la **lexicalisation** dans une deuxième langue des sens véhiculés par un mot polysémique d'une LS ne constitue qu'un des moyens de traduire ces sens. Une autre possibilité consiste à utiliser le même équivalent dans la LC pour traduire les différents sens véhiculés par le mot source dans des contextes différents dans lesquels les éléments des deux langues partagent des sens (ce qui est le cas des **ambiguïtés parallèles**). Il se peut également que des unités sémantiques de type différent soient utilisées dans la LC, ce qui implique un **changement de niveau**. Ce changement de niveau correspond à l'utilisation

mot *igluksaq* signifie « matériel pour la construction de maisons » et ne décrit donc pas un type particulier de neige. Un autre mot est le *saumavuq*, considéré comme signifiant « couvert de neige », mais qui, en réalité, signifie « il a été couvert », sans faire référence à la neige de manière concrète.

[46] D'après Whorf (1940), en anglais il n'existe qu'un mot (*snow*) pour désigner la neige qui tombe, la neige par terre, la neige entassée comme de la glace, la neige fondue et la neige emportée par le vent, tandis qu'un Eskimo utiliserait des mots différents pour ces types de neige. Pourtant, d'après Pullum (*ibid*.), en anglais on trouve le mot *snow*, quand il s'agit de flocons pelucheux et blancs, *slush* pour la neige presque fondue, *sleet* pour la neige qui tombe à moitié fondue, et *blizzard* quand elle tombe de manière dense. Ceci montre que, non seulement les généralisations concernant la langue des Eskimos ne sont pas fondées, mais aussi les remarques concernant l'anglais.

[47] Boas (*ibid*.) souligne justement qu'en anglais les notions exprimées par les Eskimos par quatre racines distinctes peuvent être exprimées par des phrases impliquant la même racine (*snow*).

de **paraphrases** et d'**additions**, qui accompagnent souvent des contextes spécifiques à une culture dans les traductions. Par conséquent, la différence entre langues ne doit pas être perçue comme relevant de ce qui « peut » et ce qui « ne peut pas être dit », mais comme ce qui est exprimé par des éléments lexicaux distincts dans une langue donnée et de manière différente dans une autre langue (Fuchs, 1996 : 86-87).

La problématique de la lexicalisation divergente des sens dans des langues différentes, nourrie par les observations des anthropologues sur les différences culturelles censées refléter des différences conceptuelles, touche de manière très claire à la problématique de la traductibilité des unités lexicales d'une langue dans une autre. Les divergences au niveau de la lexicalisation ne sont cependant pas considérées comme un obstacle insurmontable à la traduction. Von Humboldt (1816)[48] défend ainsi la possibilité de toute traduction, même s'il soutient qu'aucun mot d'une langue n'équivaut parfaitement à un mot d'une autre[49] (sauf à considérer les expressions qui désignent des objets purement corporels). Selon Sparck Jones (1986 : 24), la traduction est habituellement conçue comme la manière dont la même idée peut être exprimée par des mots différents, ce qui justifierait l'existence d'une classification conceptuelle valable pour toutes les langues. Cette classification pourrait être un thésaurus, dont les descripteurs seraient considérés comme des **classificateurs interlinguaux** ou des **catégories classificatoires**. Elle souligne, néanmoins, que ceci ne signifie pas que toutes les entrées d'un thésaurus sont interlinguales, étant donné la non représentation de certaines idées dans certaines langues. Cette non existence d'équivalents dans d'autres langues pour certains éléments linguistiques (**éléments uniques** [50]) ne constitue pas non plus un obstacle à la traduction pour Tirkkonen-Condit (2004). Ces éléments ne sont tout simplement pas manifestés de la même manière dans d'autres langues – ils n'y sont pas lexicalisés –, ce qui n'empêche pourtant pas leur traduction.

[48] *Sur la traduction*. Introduction à *l'Agamemnon*, repris dans Wilhelm Von Humboldt, *Sur le caractère national des langues et autres écrits sur le langage*, Editions du Seuil, 2000, p. 33-47.

[49] Von Humboldt parle d'une « synonymique des langues », en considérant les langues différentes comme autant de synonymes, dans le sens où chacune exprime le concept avec une nuance, avec telle ou telle connotation.

[50] Ces éléments, d'après Tirkkonen-Condit, peuvent être lexicaux, phrastiques, syntaxiques ou textuels.

La question du niveau lexical rejoint aussi, d'une certaine manière, la problématique développée par Wierzbicka autour du **métalangage sémantique naturel** (1992 : 20-21). Ce métalangage est constitué d'universaux lexicaux correspondant aux concepts humains élémentaires disponibles dans toutes les langues et dont la combinaison correcte permet l'expression de tout ce qui peut être énoncé. Par conséquent, tout ce qui peut être dit dans une langue peut être traduit, sans changement de sens, dans une autre.

Dans le domaine de la psycholinguistique, la notion du **niveau lexical** est liée à la question de la **catégorisation** au sein de langues différentes. Malt *et al*. (2003) illustrent un ensemble de relations possibles entre les catégories des langues. L'une de ces relations est celle d'**imbrication** (*nesting*), qui désigne l'existence de distinctions plus fines, dans une langue que dans une autre, à l'intérieur d'un même domaine[51]. Cette relation ne signifie pas toujours l'absence d'une telle distinction au sein de la deuxième langue car les différences peuvent y être décrites, non à l'aide de lexèmes primaires distincts mais à l'aide de phrases ou de modificateurs. Ainsi, le cas où un seul nom est utilisé pour désigner un ensemble d'objets dans une langue (par ex. *bottle* en anglais), objets désignés par des noms différents dans une autre langue, en fonction de leur matière et de leur finalité. Les différences de ce type peuvent être décrites, au sein de la première langue, à l'aide de modificateurs attachés au même nom de base (par ex. *plastic bottle, glass bottle*). Un autre exemple d'emploi d'un mot plus générique dans une langue servant à exprimer des sens exprimés par des éléments lexicaux différents dans une autre, est donné par Gale *et al*. (1993) et concerne l'ensemble des mots japonais qui peuvent être traduits en anglais par : *wearing clothes*. En japonais, il y a cinq mots différents pour *wear*, utilisés en fonction de la partie du corps impliquée. En anglais, les locuteurs ne font pas des distinctions de type "*wearing shoes*" vs. "*wearing shirt*".

Les correspondances de traduction étudiées dans ce travail sont repérées exclusivement au niveau des mots et concernent donc les deux premières catégories de correspondances décrites par Mauranen (voir le début de ce paragraphe). Les unités sémantiques se situant à d'autres niveaux que le niveau

[51] Ainsi le type de relation décrite par l'argument sur les mots en Eskimo et en anglais désignant la neige.

lexical ne seront pas prises en considération. Pourtant, si l'élimination des unités de ce type est relativement facile lors d'un processus de repérage manuel de traductions, elle devient relativement compliquée dans le cadre d'un processus d'alignement automatique[52].

Comme nous le verrons plus loin, de nombreux travaux sur la désambiguïsation admettent l'hypothèse que les sens différents d'un mot source sont lexicalisés de manière différente dans d'autres langues. Souvent, cette hypothèse est nuancée par la reconnaissance de l'existence de cas d'ambiguïté traductionnelle entre les langues, mais l'étendue de ce phénomène et son impact sur la possibilité de repérage de sens à l'aide d'équivalents de traduction et de la désambiguïsation sont rarement évalués. Avant d'exposer notre propre méthode, nous allons analyser les possibilités de correspondance entre les sens des unités lexicales polysémiques d'une LS et leurs équivalents de traduction dans une autre langue. L'analyse des correspondances possibles entre les mots de deux langues servira ainsi à mieux comprendre la nature des correspondances que nous avons établies.

2.3. Correspondances de traduction des mots polysémiques

2.3.1. Distinction complète des sens

Les différents équivalents de traduction dans une langue cible (LC) d'un mot ambigu source sont censés lexicaliser dans la deuxième langue les sens véhiculés par le mot source. Ces équivalents sont donc souvent considérés comme des indices de **repérage** et de **distinction des sens** du mot ambigu source. Cette hypothèse, bien que discutable, est sous-jacente à des méthodes qui utilisent des informations de traduction afin d'identifier les sens des mots ambigus de la LS pour la désambiguïsation (Resnik et Yarowsky, 2000 ; Dyvik, 1998a, 2003, 2005 ; Diab, 2003) et l'annotation sémantique (Diab et Resnik, 2002 ; Resnik, 2004)[53]. Dans ces méthodes, la désambiguïsation repose sur l'hypothèse

[52] Nous reviendrons sur ce sujet au paragraphe 1.3.8. du chapitre 4, qui traite des étapes de prétraitement du corpus d'apprentissage.

[53] Les principes sous-jacents à ce type de méthodes ainsi que leur fonctionnement seront décrits en détail dans le paragraphe 2 du chapitre 2.

que chaque équivalent (par ex. *a, b, c*) correspond à un sens précis du mot ambigu source (par ex. *A*) ; hypothèse pourtant nuancée si l'on considère les ambiguïtés parallèles entre langues.

Pour que cette hypothèse soit vraie dans son acception extrême, il faudrait que le nombre d'équivalents de traduction (*N*) soit égal au nombre de sens (*n*)[54] véhiculés par l'unité lexicale source (*n*=*N*). Autrement dit, l'espace sémantique occupé par une unité source serait considéré comme étant découpé et couvert par des unités lexicales distinctes dans la LC, dont chacune représenterait un sens différent pertinent de l'unité source. Dans ce cas, les équivalents illustreraient l'ensemble des sens exprimés par l'unité de la LS et pourraient, ainsi, servir à la désambiguïser de manière complète. Ce type de relation est décrit schématiquement par la figure 3, où l'espace sémantique couvert par une unité source (A)[55] est découpée en des régions sémantiques plus petites, dont le contenu est exprimé par des équivalents différents (*a*, *b*, *c*) dans la LC.

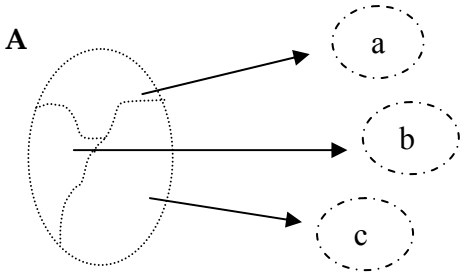


Figure 3. Equivalents lexicalisant dans la LC les sens de l'unité polysémique de la LS

[54] Il faut souligner qu'il s'agit ici d'une représentation schématique et abstraite des correspondances inter-langues qui peuvent exister au niveau sémantique et des possibilités d'utilisation des relations de traduction comme indices pour le repérage de sens et la désambiguïsation. Le fait que nous faisons référence au nombre de sens des mots polysémiques ne signifie pas que nous les avons repérés à l'avance ni que nous en connaissons le nombre. Nous utilisons ce moyen comme un simple appareil de représentation des distinctions possibles au niveau sémantique et des possibilités de correspondance entre les mots de deux langues. La même remarque est valable pour les paragraphes suivants, où nous faisons référence au nombre de sens des mots polysémiques source.

[55] Nous décrivons ici l'espace sémantique couvert par un mot avec des frontières bien précises, par convenance. Bien évidemment, cet espace peut évoluer, s'étendre ou se restreindre dans le temps, et cette évolution dépend de paramètres linguistiques (comme la catégorie grammaticale du mot), et de facteurs extra-linguistiques (comme l'appartenance du mot à la langue générale ou à un domaine de connaissances restreint et spécialisé).

L'hypothèse sous-jacente à ces méthodes de repérage de sens postule que la mise en correspondance du mot ambigu avec chacun de ses équivalents permettrait la distinction des sens véhiculés par le mot source. Ainsi la formation de paires de type '*A_a*', '*A_b*', '*A_c*', etc. rendrait évidentes les distinctions sémantiques « cachées » dans le mot source et « révélées » par les équivalents de traduction. Cette mise en correspondance pourrait donc contribuer au repérage des sens différents des mots source, ainsi qu'à leur annotation sémantique. Dans ce cas, les étiquettes des occurrences des mots source dans les textes correspondraient aux paires formées de la manière décrite ci-dessus.

Dyvik (2003) donne comme exemple de ce processus l'adjectif norvégien *lekker* et ses deux équivalents de traduction possibles en anglais : *delicious* et *pretty*. Ces équivalents indiquent une manière, parmi d'autres, de diviser la potentialité sémantique de l'adjectif norvégien. Ils sont liés à des aspects différents ou à des sous-sens reliés du mot *lekker*, qui correspondraient à des paires comme [*lekker*|*delicious*] et [*lekker*|*pretty*], ou à des ensembles d'éléments plus grands, dans le cas où plusieurs langues interviennent[56]. Dans une approche traductionnelle de la sémantique, ces ensembles pourraient être considérés comme des primitives de descriptions sémantiques. Ainsi une paire telle que [*lekker*|*delicious*] pourrait être traitée comme un type de trait sémantique attribuable à des unités lexicales et utilisable pour leur classification. De cette manière, les mots qui n'ont pas de relations biunivoques avec leurs équivalents peuvent être décomposés en éléments primitifs, capables de décrire les alternatives de traduction.

Cette hypothèse de lexicalisation des sens se révèle valide surtout dans les cas d'homonymes, où les distinctions de sens sont évidentes et généralement lexicalisées dans d'autres langues. Un exemple, régulièrement cité, illustrant la possibilité de distinguer les sens d'un homonyme par le biais de ses équivalents de traduction est le cas du mot anglais *bank* qui possède les deux sens d'« institution financière » et de « rive ». Les candidats de traduction de ce

[56] La généralisation, dans le cas de plusieurs langues, n'est pas évidente, en raison des divergences observées au niveau des distinctions d'une langue à l'autre, divergences dépendant de facteurs que nous analyserons plus loin. Nous avons déjà vu que Resnik et Yarowsky (1997) proposent, pour les travaux qui impliquent plusieurs langues, la définition d'un nombre minimum de langues dans lesquelles les distinctions doivent être retrouvées.

mot en français correspondent aux deux noms *banque* et *rive*, qui traduisent justement les deux sens du mot source. Les deux sens du mot anglais sont donc mis en évidence par ses candidats de traduction en français et la traduction d'une nouvelle occurrence de ce mot par l'un de ces équivalents montre quel est le sens véhiculé par cette occurrence (Resnik, 2004). Par exemple, si une occurrence de *bank* dans un texte est traduite par l'équivalent *banque*, il est évident que le mot est employé dans son sens financier.

Cependant, même dans le cas des homonymes, les choses sont loin d'être simples. Comme nous l'avons montré dans le paragraphe sur les relations entre les sens des unités polysémiques (§1.2.), il peut arriver qu'un mot soit caractérisé, en même temps, par homonymie et polysémie (Fuchs, 1996 : 10 ; Pustejovsky, 1995 : 27). Ainsi, il est possible que l'un des deux homonymes véhicule d'autres sens, qui sont des manifestations d'un même sens principal lorsqu'il apparaît dans des contextes différents. Dans le cas de *bank*, les sens « institution financière » et « bâtiment » peuvent être véhiculés par l'homonyme qui véhicule le sens financier. Des distinctions de ce type, quoique pertinentes dans la LS – et probablement dans la LC – sont plus rarement lexicalisées au sein de la deuxième langue. Les informations de traduction peuvent donc ne pas être suffisantes pour distinguer les sens lors de tels cas d'ambiguïté.

Dans le paragraphe suivant, nous allons aborder plus précisément la question de la non lexicalisation des sens dans la langue cible et de ses conséquences sur la désambiguïsation.

2.3.2. Recouvrement de sens partiel ou total

Le phénomène de **recouvrement de sens** entre unités lexicales de deux langues est souvent désigné dans la littérature par le terme d'**ambiguïté traductionnelle** ou d'**ambiguïté parallèle**. Le recouvrement de sens entre un mot source et son équivalent de traduction consiste en ce que l'équivalent présente la même ambiguïté dans la LC ou, du moins, une ambiguïté similaire coïncidant en partie avec celle du mot source. Le recouvrement peut, par conséquent, être total ou partiel.

La relation de recouvrement total correspond au plus haut degré d'ambiguïté parallèle, cas où les mots des deux langues en relation de traduction possèdent les mêmes extensions de sens. Altenberg et Granger (2002b) préfèrent parler de recouvrement de sens lorsque les mots des deux langues présentent approximativement les mêmes extensions de sens, et de « divergence de sens », dans le cas où les mots ont des extensions de sens différentes, ce qui correspondrait au cas de recouvrement partiel[57]. Dans les deux cas de recouvrement de sens, partiel ou total, l'équivalent traduit dans la LC plus d'un sens du mot source ; l'ambiguïté du mot source est donc, d'une certaine manière, « préservée » dans la LC[58]. Les mots ambigus source qui entrent dans ce type de relation avec leurs équivalents ne sont pas, d'après Salkie (2002c), ambigus du point de vue de la traduction[59].

L'apparition et la fréquence des cas d'ambiguïté traductionnelle dépendent des langues source et cible. De tels cas sont le plus souvent observés entre langues proches, où il est davantage probable de rencontrer des éléments présentant une ambiguïté similaire[60]. Ainsi, l'ambiguïté d'un mot source au sein de la LS ne signifie pas nécessairement que ce mot est toujours ambigu du point de vue de la traduction ; cela dépend fortement de la LC. La préservation de la polysémie des mots source dans la LC rend inutile la désambiguïsation des mots source lors de la traduction. Néanmoins, cette préservation rend compte du fait que l'ambiguïté traductionnelle constitue un obstacle à l'utilisation des équivalents de traduction d'un mot comme indices de distinction de ses sens. Pour l'acquisition de sens, dans ces cas, une troisième langue serait nécessaire,

[57] Altenberg et Granger distinguent par ailleurs un autre type de relation entre les mots polysémiques des langues, la relation de non correspondance, où un élément n'a pas d'équivalent évident dans l'autre langue.

[58] Un exemple largement employé dans la littérature pour illustrer ce type d'ambiguïté concerne le mot français *intérêt* et son équivalent en anglais *interest*, qui peut être utilisé pour traduire les deux sens du mot français. Néanmoins, il ne faut pas exclure la possibilité de trouver d'autres équivalents de traduction pour le mot *intérêt* lorsqu'il est utilisé en contexte, par exemple au sein d'un corpus parallèle.

[59] Au contraire, un mot ambigu source dont les équivalents correspondent à des sens différents est considéré comme ambigu du point de vue de la traduction. Dans ce cas, le traducteur doit en effet distinguer entre les équivalents qui correspondent à des sens différents. Ces informations peuvent aider à la distinction des sens des mots source.

[60] Nous analyserons plus en détail l'impact du facteur de la distance inter-langue sur l'apparition d'ambiguïtés traductionnelles dans le paragraphe 2.3.2 du chapitre 2.

dans laquelle la polysémie ne serait pas préservée et dont les équivalents constitueraient des indices de distinctions des sens.

L'ambiguïté parallèle peut exister entre les mots de deux langues à des degrés variés. Un équivalent peut « englober » tous les sens (*n*)[61] exprimés par le mot source ; dans ce cas, la relation entre sens et équivalent peut être décrite comme *n:1*. Ce cas peut être caractérisé comme **ambiguïté traductionnelle totale** ou **recouvrement de sens total** et peut être considéré comme l'un des extrêmes d'un **continuum d'ambiguïté traductionnelle**. Ce continuum peut être décrit par un intervalle semi-ouvert (ouvert à gauche) dont l'extrémité gauche désignerait les cas de distinction complète des sens (où le nombre des équivalents est égal au nombre des sens : *n*=N), et qui ne sont donc pas des cas d'ambiguïté traductionnelle[62].

En reprenant la terminologie de Salkie (2002c), l'extrémité gauche du continuum désignerait les mots polysémiques qui ne sont pas ambigus du point de vue de la traduction et l'extrémité droite, ceux qui le sont. A l'intérieur de ce continuum, nous pourrions trouver les cas intermédiaires de correspondance en matière d'ambiguïté traductionnelle, entre les mots ambigus d'une langue et leurs équivalents de traduction dans une autre. Ces cas intermédiaires concerneraient les autres possibilités de correspondance entre sens lexicaux et équivalents, où certains sens véhiculés par le mot polysémique sont traduits par des éléments distincts dans la LC et d'autres exprimés par un mot polysémique ou plus générique dans la LC. Dans ce cas, le nombre de candidats de traduction du mot source serait supérieur à un (*N*>1) et inférieur à celui des sens véhiculés par le mot source (*N*<*n*).

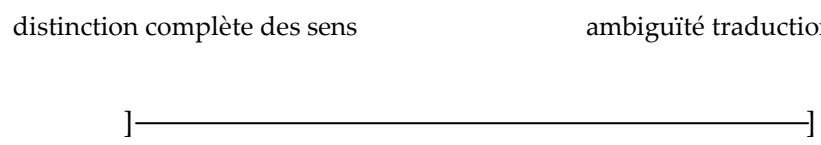


Figure 4. Intervalle semi-ouvert d'ambiguïté traductionnelle

[61] Voir note de bas de page n° 54, sur la référence au nombre de sens de l'unité polysémique source.
[62] C'est le cas que nous avons décrit dans le paragraphe précédent, où le nombre des équivalents (*N*) est égal au nombre des sens (*n*), et la relation sens-équivalents est de type *n* : *N*.

La fréquence d'observation de correspondances intermédiaires dans des textes est supérieure à celle correspondant aux deux extrêmes du continuum d'ambiguïté. Cependant, l'éventuel apport des correspondances intermédiaires pour l'acquisition des sens est plus difficilement mesurable que dans les cas extrêmes. Cet apport est en effet variable et dépend fortement de la nature des sens exprimés par les équivalents, ainsi que des relations éventuellement entretenues entre les sens. Ainsi, il se peut que, même au niveau du même mot source, certains des équivalents lexicalisent des sens du mot source dans la LC et qu'ils correspondent, par conséquent, de manière évidente à des distinctions sémantiques propres à ce mot ; ce qui peut ne pas être le cas pour d'autres équivalents. Un exemple de ce type de relation est illustré dans la figure 5 : dans le cas d'un mot source qui véhicule un nombre de sens différents, certains sens peuvent être lexicalisés par des équivalents différents (*a*, *b*), tandis que les autres peuvent être exprimés par le même équivalent de traduction (c), probablement polysémique dans la LC.

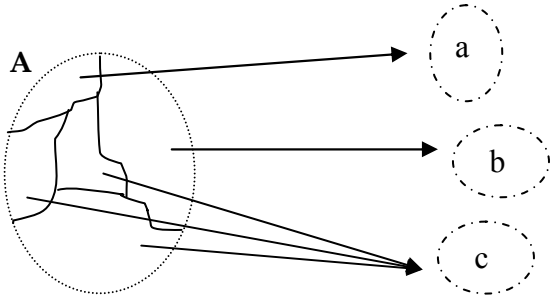


Figure 5. Cas intermédiaire entre l'ambiguïté traductionnelle et la distinction complète des sens

Les équivalents de traduction peuvent donc être également ambigus et exprimer l'ensemble des sens véhiculés par le mot source, comme nous l'avons montré plus haut, ou une partie de ces sens. La description de ces cas de correspondance entre sens des mots source et cible ne doit pas exclure d'une telle analyse les cas de mots source dont les sens ne peuvent être clairement distingués, comme les mots vagues. Le sens de ces mots est plutôt déterminé par le contexte dans lequel ils apparaissent, ce qui empêche l'identification de correspondances claires entre ces sens et les mots pouvant les traduire dans la LC.

Il est à noter que la sémantique des équivalents peut diverger de manière importante de celle du mot source. Ils peuvent donc exprimer des sens qui ne sont pas exprimés par le mot source. Dans ce cas, l'ensemble des sens des équivalents pourrait être mis en évidence par l'inversion de la direction de traduction et par le repérage de leurs propres équivalents de traduction possibles dans la LS[63].

Les unités lexicales de deux langues qui entretiennent une relation de recouvrement partiel ont des extensions de sens divergentes. Les correspondances de ce type entre les unités lexicales de deux langues se prêtent à une description en des termes « topologiques », comme dans le modèle proposé par Victorri et Fuchs (1996). Ainsi, on pourrait dire qu'une partie de l'**espace sémantique** occupé par le mot polysémique de la LS – qui peut correspondre à un ou plusieurs sens de ce mot – est occupée dans la LC par un seul mot, tandis que le reste de cet espace peut être couvert par d'autres mots de la LC. Du côté de la LC, un mot peut également véhiculer des sens multiples, éventuellement différents de ceux exprimés par le mot source. Si la direction de traduction était inversée, une partie de son espace sémantique – qui correspond à un ou plusieurs sens – serait couverte par le mot source tandis que le reste de cet espace serait couvert par d'autres unités lexicales de la LS. Dans ces cas, les mots des deux langues pourraient être caractérisés comme « pareillement ambigus » quant à certains de leurs sens, mais non pour la totalité.

La figure 6 illustre schématiquement l'ensemble des correspondances de traduction extraites de notre corpus d'apprentissage pour le mot anglais *plant*. Ce mot a six équivalents de traduction grecs au sein du bitexte : *φυτό* (fyto), *εργοστάσιο* (ergostasio), *εγκατάσταση* (egkatastasi), *μονάδα* (monada) et *σταθμός* (stathmos). En inversant la direction de traduction, nous constatons que chacun des mots grecs entretient des relations de traduction avec un ensemble de mots anglais parmi lesquels se trouve, bien entendu, le mot *plant*. Quelques-unes de ces correspondances de traduction mettent en évidence des relations de recouvrement de sens partiel entre le mot source (*plant*) et son équivalent. En effet, le repérage d'un grand nombre de correspondances de traduction pour un

[63] Ainsi la méthode des Miroirs Sémantiques, qui désambiguïse les mots des deux langues impliquées (cf. chapitre 9).

équivalent de traduction du mot source, lors de l'inversion de la direction de traduction, ne suffit pas à lui seul à garantir un cas de recouvrement de sens partiel.

Plus concrètement, dans notre exemple, il se peut que quelques-uns des (ou tous les) équivalents anglais d'un mot grec soient (plus ou moins) synonymes et qu'ils n'indiquent pas, par conséquent, des distinctions sémantiques du mot.

Il se peut également que, dans certains cas, les mots anglais mis en correspondance avec les mots grecs soient des synonymes du mot source anglais. Tel est le cas des mots *factory*, *industry* et *station* qui traduisent le mot *εργοστάσιο* au sein du bitexte : ces équivalents sont des synonymes entre eux et, également, des synonymes du mot source *plant*. Les correspondances de traduction de *εργοστάσιο* montrent donc qu'il ne véhicule pas, dans le corpus, des sens non véhiculés par le mot *plant*. En revanche, le mot *plant* véhicule des sens non exprimés par l'équivalent *εργοστάσιο*, à savoir son sens « végétal ».

Il arrive, dans certains autres cas, que les traductions des mots grecs révèlent des sens qui n'ont pas de relation avec les sens du mot source. Tel est, par exemple, le cas des mots *point*, *centre*, *stop* et *shelter*, équivalents de traduction du mot *σταθμός* (stathmos) dans le corpus. Ces équivalents traduisent les instances du mot *σταθμός* qui véhiculent le sens d' « arrêt » et d' « abri », sens qui ne sont pas véhiculés par *plant*. Dans ce cas, la sémantique de l'équivalent *σταθμός* diverge de celle du mot source dans la mesure où il exprime des sens qui ne sont pas exprimés par le mot source. L'inversion de la direction de traduction et le repérage des équivalents de traduction des équivalents d'un mot source permettent donc une analyse de la sémantique des équivalents de traduction des mots source, ce qui rend possible l'établissement de correspondances plus fiables entre les espaces sémantiques des mots des deux langues.

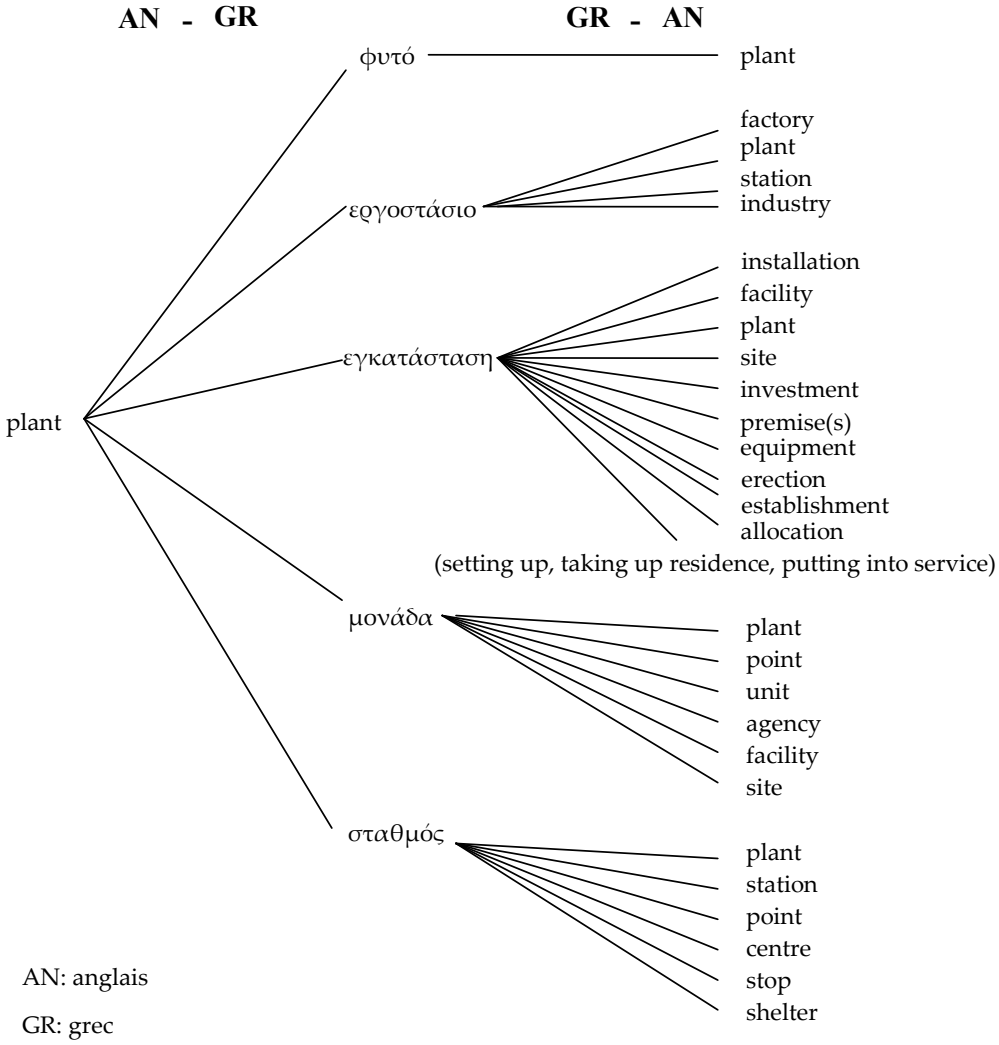


Figure 6. Correspondances de traduction pour *plant* et inversion de la direction de traduction

La **distance typologique** entre deux langues constitue un paramètre qui influence beaucoup l'existence de relations d'ambiguïté parallèle et de recouvrement entre les unités lexicales en relation de traduction. Cette distance entre les langues constitue donc un paramètre qui conditionne la possibilité d'utilisation des équivalents de traduction en tant qu'indices pour la désambiguïsation des unités lexicales de la LS. Nous reviendrons sur ce sujet dans le §2.3.2. du chapitre 2, où nous analyserons les paramètres qui influencent le fonctionnement des méthodes traductionnelles d'acquisition de sens.

En prenant en compte la sémantique des mots de la LC, nous pouvons distinguer deux cas de recouvrement total :

- celui où l'équivalent exprime tous les sens du mot source, sans exprimer de sens qui ne sont pas exprimés par ce mot
- celui où l'équivalent exprime tous les sens du mot source, plus d'autres sens qui lui sont propres.

En revanche, la relation de recouvrement partiel peut servir à décrire trois cas :

le cas où l'équivalent exprime seulement une partie des sens véhiculés par le mot source, sans exprimer de sens propres (dans ce cas, le recouvrement partiel concerne la sémantique du mot source)

le cas où l'équivalent exprime tous les sens du mot source, plus d'autres qui lui sont propres (dans ce cas, le recouvrement partiel concerne la sémantique de l'équivalent)

le cas où les mots des deux langues ont des extensions de sens divergentes et où l'équivalent exprime une partie des sens véhiculés par le mot source et d'autres qui lui sont propres (dans ce cas, le recouvrement partiel concerne la sémantique des mots des deux langues)

L'établissement de l'ensemble des correspondances sémantiques possibles entre les mots de deux langues nécessiterait une analyse de la sémantique des équivalents de traduction. Cette analyse permettrait d'avoir une image claire du découpage sémantique et de la lexicalisation de sens effectués au sein de ces langues. Il faut souligner que la complexité apparente du réseau de correspondances entre les espaces sémantiques ne pose néanmoins pas de problème lors des tentatives d'acquisition de sens, lorsque celles-ci se focalisent sur les mots d'une seule langue. Dans ce cas, l'analyse complète de la sémantique des unités des deux langues n'est pas nécessaire. Ceci s'explique par le fait que lorsque les équivalents de traduction d'un mot dans une autre langue sont utilisés pour le repérage des sens d'un mot source, leur sémantique est « restreinte » aux sens liés à ceux du mot source. Dans une étude basée sur corpus, une telle restriction est opérée facilement, en considérant seulement les instances des équivalents qui traduisent des instances du mot source et en

ignorant les autres. Ainsi, lorsque la polysémie des équivalents diffère de celle de l'unité source, le reste de leur polysémie, qui n'est pas mis en correspondance, peut être ignoré.

2.3.3. Correspondances lexicales inter-langues au sein des corpus de traduction

L'équivalence complète entre les mots de langues différentes est très rare en raison du découpage divergent des champs sémantiques concernés. Par conséquent, les mots traités comme équivalents de traduction dans des dictionnaires bilingues ont généralement des éventails de sens différents (Altenberg et Granger, 2002b). L'ambiguïté divergente et les extensions sémantiques différentes des équivalents de traduction fournis dans les dictionnaires expliquent que ces équivalents aient rarement la même distribution dans des textes réels et que leur degré de correspondance mutuelle soit très faible (Salkie, 1997; Viberg, 2002). Dickens et Salkie (1996) décrivent par **équivalence de traduction élémentaire** les équivalences fournies dans les dictionnaires bilingues et par **équivalence de traduction riche** les équivalences observées au sein de corpus parallèles[64]. Cette différence est due, d'après Dickens et Salkie, à la tendance des dictionnaires à ne donner qu'une image limitée de l'éventail complet des stratégies utilisées par les traducteurs expérimentés, mis en évidence par une étude sur corpus.

Le faible degré de correspondance mutuelle des équivalents fournis par les dictionnaires dans des textes réels peut exprimer l'existence de différences au niveau du système de la langue et de manques lexicaux dans la LS ou dans la LC, mais peut aussi être dû à des différences stylistiques ou fonctionnelles (Altenberg et Granger, 2002b). L'absence d'équivalent de traduction évident pour une unité

[64] Les lexèmes de la LS qui apparaissent fréquemment dans le corpus et qui sont traduits par un grand nombre d'équivalents différents (et présentent donc une équivalence de traduction riche) mériteraient un traitement spécial dans les dictionnaires bilingues (Dickens et Salkie, 1996) : ils pourraient être étiquetés de manière à inciter l'utilisateur à considérer un éventail de traductions plus riche que celui fourni par le dictionnaire ; l'utilisateur pourrait être renvoyé à un dictionnaire monolingue contenant une image plus complète des possibilités de traduction ; des informations concernant les stratégies possibles pour la traduction du lexème en question pourraient être fournies, en donnant davantage d'exemples et en indiquant comment ceux-ci pourraient servir à sélectionner la traduction adéquate en contexte.

lexicale source pose souvent des difficultés au traducteur, et peut provoquer soit sa non traduction soit un large éventail de traductions. Le deuxième cas témoigne de l'effort du traducteur pour rendre le mot source dans la LC, en recourant à des manières diverses en fonction du contexte.

Les **corpus de traduction** (ou **corpus parallèles**), qui sont constitués de textes originaux d'une langue et de leurs traductions dans une autre langue, offrent la possibilité d'explorer toute une gamme de types de traduction, qui reflètent les degrés différents de correspondance inter-langue. Dans ces corpus, on trouve, d'une part, des éléments qui ont des équivalents « attendus » fortement récurrents et, d'autre part, des éléments ayant une variété très grande de traductions « inattendues », souvent difficiles à classifier. Salkie (2002c) décrit les correspondances par un **spectre** qui va des éléments **systématiques** du point de vue de la traduction aux éléments **asystématiques**, traduits différemment à chaque occurrence. Le traitement automatique – mais également humain – des éléments asystématiques et des cas intermédiaires de ce continuum s'avère évidemment beaucoup plus difficile que celui des éléments systématiques.

Les traductions inattendues trouvées dans les textes sont souvent considérées comme des produits de la performance du traducteur. Néanmoins, des raisons expliquent souvent ces choix. Dans le cas d'un mot source vague, par exemple, où le sens d'une instance du mot est fortement dépendant du contexte, il se peut que le traducteur soit obligé d'inventer de nouvelles solutions pour traduire ses différentes occurrences. Les corpus de traduction permettent de découvrir les motifs sous-jacents à l'inventivité du traducteur et aux traductions les plus inattendues trouvées dans les textes.

La systématicité traductionnelle est considérée par Salkie (*ibid*.) comme une relation entre deux **systèmes linguistiques**, et la non-systematicité comme une relation au niveau de la **pratique textuelle**. Néanmoins, les différences systématiques observées au niveau de la pratique textuelle – telle qu'elle peut être constatée dans un corpus de traduction – peuvent refléter des différences au niveau des systèmes correspondants. Salkie souligne le besoin de prendre en compte les différences au niveau de la pratique textuelle entre les langues en proposant l'inclusion de règles de fréquence au niveau du système. Cette proposition poserait pourtant un ensemble de problèmes, étant donné que les

connaissances pouvant être tirées des corpus sur la systématicité des mots polysémiques sont soumises aux contraintes inhérentes aux études basées sur corpus, autrement dit, elles dépendent fortement de la taille et de la constitution des corpus utilisés.

CONCLUSION

Au cours de ce chapitre, nous avons présenté un ensemble de facteurs linguistiques et extra-linguistiques qui expliquent pourquoi une réponse unanime à propos du repérage et de la délimitation des sens lexicaux est impossible. La variabilité des relations entre les sens est telle que même l'établissement d'une typologie et la catégorisation des mots polysémiques par rapport à des types prédéfinis sont difficiles. Une telle caractérisation des mots polysémiques impliquerait l'analyse des relations entretenues entre les sens, aspect souvent négligé dans la pratique lexicographique contemporaine. Cette question se complique encore davantage dans le cadre du traitement de la polysémie lexicale dans un contexte bilingue, où la complexité de l'analyse augmente relativement au besoin de création de correspondances sémantiques entre les unités lexicales des deux langues.

Dans le deuxième chapitre, nous allons présenter un ensemble de méthodes automatiques qui ont été proposées, dans un cadre monolingue et bilingue, pour le repérage des sens des unités lexicales polysémiques. Nous montrerons ensuite comment la résolution de la polysémie peut être effectuée dans un cadre automatique, par référence à des ressources sémantiques préétablies ou à des résultats de méthodes de repérage automatique des sens.

CHAPITRE 2

REPERAGE AUTOMATIQUE DE SENS LEXICAUX

INTRODUCTION

La remise en question de l'adéquation des ressources lexicales et sémantiques prédéfinies au traitement automatique a généré l'émergence de méthodes de repérage (ou d'acquisition) automatique de sens. Il s'agit de méthodes **dirigées par les données** contenues dans des corpus textuels monolingues et bi- (ou multi-) lingues. L'objectif de ces méthodes empiriques est l'identification des sens véhiculés par les mots ambigus d'une langue. Elles permettent l'acquisition d'inventaires de sens pertinents pour les données traitées mais, surtout, facilement exploitables dans un cadre de traitement automatique. En outre, étant fondées sur des traitements statistiques, ces méthodes sont opératoires dans le cadre de langues différentes.

Les inventaires sémantiques générés de cette manière peuvent constituer les ressources sémantiques requises par les méthodes de désambiguïsation lexicale. Ils fournissent les sens possibles (ou **sens candidats**) des mots, à partir desquels

un doit être sélectionné comme décrivant le sens d'une instance d'un mot ambigu en contexte. Ce chapitre sera donc consacré à une présentation des méthodes automatiques d'acquisition de sens, tandis que le chapitre suivant traitera la question de la désambiguïsation.

1. Repérage automatique de sens dans un cadre monolingue

1.1. Méthodes dirigées par les données

1.1.1. Apprentissage basé sur des régularités distributionnelles

L'hypothèse sous-jacente aux méthodes monolingues de repérage de sens est l'**hypothèse distributionnelle** du sens (Harris, 1954), selon laquelle les différents **sens** des mots sont reflétés dans leurs **usages** au sein des textes. Les sens sont repérés à un niveau d'abstraction supérieur à celui des occurrences. Pour atteindre ce niveau, les usages présentant certaines **régularités** sont regroupés. Les groupes générés sont supposés décrire les différents sens des mots en question. La discrimination des sens lexicaux est ainsi souvent réduite au problème de repérage de classes (ou de clusters) de contextes similaires telles que chaque classe représente un sens. Etant donné un mot polysémique utilisé dans un ensemble de contextes différents, le processus d'acquisition des sens consiste à regrouper les instances du mot, en déterminant ses contextes qui présentent la plus grande similarité entre eux.

Le repérage automatique des sens lexicaux se fonde donc sur les données trouvées dans les textes ; les méthodes développées dans ce but sont alors des **méthodes dirigées par les données** (*corpus driven*). Ces méthodes ne présupposent pas l'utilisation de ressources prédéfinies (dictionnaires, thésaurus ou ontologies). Les distinctions sémantiques qui caractérisent un mot sont identifiées sur la base de données textuelles, au cours d'une étape de traitement qui ressemble à l'**apprentissage automatique** (*machine learning*). Lors de cette étape, les informations pertinentes pour l'étude de la sémantique des mots sont repérées dans les textes et analysées. Les résultats de cette analyse rendent

évidents les sens véhiculés par les mots, qui peuvent, par la suite, être modélisés. Les descriptions sémantiques engendrées sont généralement associées aux informations contextuelles qui servent au repérage et à la distinction des sens et peuvent ensuite être exploitées par des méthodes de désambiguïsation lexicale pour la sélection du sens de nouvelles occurrences des mots polysémiques.

1.1.2. Apprentissage non supervisé

Les techniques d'apprentissage automatique utilisées pour l'identification des sens lexicaux sont **non supervisées**. L'apprentissage non-supervisé ne présuppose pas une sortie définie *a priori*. Les algorithmes utilisés apprennent des motifs à partir des paramètres d'entrée, sans tenter d'établir des correspondances avec des catégories spécifiées à l'avance, comme c'est le cas dans l'apprentissage supervisé. De tels algorithmes sont les algorithmes de **clustering**, qui permettent la classification d'un ensemble d'objets dans des groupes différents ou, autrement dit, le partitionnement d'un ensemble de données en sous-ensembles (clusters). Le partitionnement se fait de telle manière que les données d'un sous-ensemble partagent quelque(s) trait(s) commun(s). Le clustering se base sur la proximité des objets par rapport à une **mesure de distance** prédéfinie[65]. Cette mesure constitue un paramètre très important pour le clustering. Elle détermine la manière dont la similarité ente deux éléments est calculée et influence, par conséquent, le contenu des clusters obtenus.

L'apprentissage non supervisé appliqué à l'acquisition de sens consiste à regrouper les instances sémantiquement similaires de mots, sur la base de l'hypothèse du comportement distributionnel similaire des instances en question. D'après cette hypothèse, les traits provenant du contexte lexical d'une instance d'un mot la caractérisent (par ex. ses cooccurrents) et la similarité des instances est calculée en termes de similarité des ensembles de **traits contextuels** correspondants. Le regroupement des instances s'effectue à l'aide de techniques de clustering (Pedersen et Bruce, 1997a ; Widdows et Dorow, 2002). La représentation des traits d'un mot peut aussi prendre la forme d'un **vecteur**

[65] Certains éléments peuvent être considérés comme proches entre eux, relativement à un aspect, et plus distants par rapport à un autre.

(comme dans la recherche d'information) et le **calcul de similarité** peut correspondre au **calcul de la distance** dans un espace multidimensionnel (Schütze, 1992, 1998 ; Pantel et Lin, 2002, 2003 ; Purandare et Pedersen, 2004a,b). Etant donné une métrique de distance, les vecteurs peuvent être clusterisés, ce qui permet ensuite la formation de classes de mots[66]. Dans une tâche d'acquisition de sens, le nombre de clusters possibles n'est pas spécifié à l'avance, ni les étiquettes de chaque cluster. Cette absence d'étiquettes prédéfinies caractérisant les clusters désigne cette tâche davantage comme une tâche de **discrimination** que comme une tâche d'**identification** de sens (Pedersen, 2007). L'identification des sens obtenus et l'attribution d'étiquettes spécifiques pourraient constituer une prochaine étape de traitement[67].

1.1.3. Algorithmes de clustering

1.1.3.1. Types d'algorithmes

Les algorithmes utilisés pour le clustering des instances sont de deux types : les **algorithmes hiérarchiques** et les **algorithmes de partitionnement**. L'algorithme constitue la stratégie de recherche qui définit la manière dont les instances seront traitées. La sélection des clusters qui sont divisés ou fusionnés à chaque itération de l'algorithme s'opère en fonction d'un critère spécifié à l'avance[68]. La différence principale entre les deux types d'algorithmes précités est

[66] D'après Resnik (1995), l'interprétation de la métrique de distance utilisée constitue une difficulté pour la plupart des méthodes distributionnelles, dans le sens où les classes de mots résultant du clustering distributionnel, caractérisées habituellement comme « sémantiques », décrivent bien souvent des facteurs syntaxiques, pragmatiques ou stylistiques.

[67] Cette étape pourrait impliquer, par exemple, la mise en correspondance des clusters construits pour un mot avec les sens proposés pour le mot dans un dictionnaire ou une autre ressource lexicale. Cette correspondance pourrait être établie par un humain ou automatiquement, à l'aide de métriques de similarité, qui permettraient de caractériser les sens obtenus en fonction de ceux décrits au sein de la ressource.

[68] Ce critère définit la manière dont l'algorithme calcule la similarité entre clusters. Dans les méthodes qui utilisent le « lien simple », la similarité entre deux clusters correspond à la similarité entre leurs membres les plus similaires, c'est-à-dire à la distance minimale entre leurs éléments. Les clusters les plus proches sont ainsi combinés. Par contre, les méthodes qui utilisent le « lien complet » exploitent la similarité entre les membres des clusters qui sont le moins similaires ; la similarité entre deux clusters correspond, donc, à la distance maximale entre leurs éléments. Dans ce cas, ce sont les clusters les plus éloignés qui sont combinés. Dans les méthodes qui utilisent le

que les algorithmes hiérarchiques forment des clusters successivement, en utilisant les clusters précédemment établis, tandis que les algorithmes de partitionnement déterminent tous les clusters en même temps.

1.1.3.2. Spécifications sur les algorithmes hiérarchiques

Une distinction plus fine peut pourtant être établie au sein de la catégorie des algorithmes hiérarchiques, qui se divisent en **algorithmes d'agglomération** (*bottom-up*, en anglais) et **de division** (*top-down*, en anglais). Ces deux types d'algorithmes procèdent de manière itérative, les premiers en fusionnant et les deuxièmes en divisant des clusters à chaque étape. Les algorithmes d'agglomération placent au préalable chaque instance dans un cluster séparé et fusionnent ensuite une paire de clusters à chaque itération, formant ainsi des clusters de plus en plus grands, jusqu'à ce qu'il n'en reste plus qu'un. Les algorithmes hiérarchiques de division commencent en plaçant toutes les instances dans le même cluster, puis en le divisant en deux à chaque itération, jusqu'à ce que chaque instance se retrouve dans un cluster distinct.

Les différentes étapes d'un algorithme hiérarchique peuvent être représentées à l'aide d'un arbre, appelé **dendrogramme**. Un des problèmes liés à ce type d'algorithmes est la définition d'une **coupe** d'arbre, qui détermine à quel point l'agglomération ou la division des sens doit s'arrêter. Le seuil de la coupe, c'est-à-dire la hauteur à laquelle celle-ci est située dans l'arbre, est important, car il conditionne la précision du clustering effectué et détermine le nombre de clusters fournis. Etant donné la difficulté à définir le seuil, il arrive, dans certains travaux utilisant ce type d'algorithmes, que l'ensemble de l'arbre (représentant l'historique des fusions ou des scissions) soit fourni, en laissant la décision finale à l'utilisateur humain[69].

Un autre inconvénient inhérent à cette approche est l'impossibilité de **chevauchement** entre clusters. Une donnée peut être proche de données appartenant à deux clusters différents, mais la représentation de cette proximité

« lien moyen », la similarité correspond à la similarité moyenne entre toutes les paires d'éléments des clusters ; les plus proches sont combinés.

[69] Ceci est le cas, par exemple, dans le travail de Kaji (2003).

n'est pas possible avec l'utilisation d'un algorithme hiérarchique. Ainsi, dans le cas de l'acquisition de sens, si un trait contextuel est pertinent pour le repérage de plusieurs sens, il doit être lié à des clusters différents, mais ce chevauchement de clusters ne peut être représenté à l'aide d'algorithmes de ce type.

1.1.3.3. Spécifications sur les algorithmes de partitionnement

L'autre type d'algorithmes de clustering, les algorithmes de partitionnement, divisent un ensemble d'instances en un **nombre prédéterminé de clusters** sans passer par toute la série de comparaisons entre paires de clusters. L'avantage principal de ces méthodes est qu'elles sont plus simples et plus rapides que les algorithmes hiérarchiques, ce qui leur permet de bien fonctionner sur de grands ensembles de données. Leurs inconvénients majeurs sont qu'elles nécessitent la définition, à l'avance, du nombre de clusters final[70] et qu'elles ne donnent pas le même résultat à chaque utilisation. Ceci est dû au fait que l'attribution initiale des instances aux clusters s'opère de manière aléatoire, ce qui signifie qu'elle peut être différente à chaque utilisation.

L'utilisation d'algorithmes de partitionnement dans une tâche de repérage de sens n'est pas évidente. La nécessité de définir le nombre des clusters *a priori* est contradictoire avec la nature de la tâche en question, où le nombre de sens des mots étudiés n'est pas connu à l'avance mais doit être justement découvert à partir des données.

[70] Par exemple, l'algorithme « k-means » (Jain *et al*., 1999) génère *k* clusters différents qui ne se chevauchent pas. La première étape du fonctionnement de cet algorithme consiste à déterminer le nombre des clusters (*k*). Les clusters sont générés de manière aléatoire et leurs centroïdes sont déterminées. Par la suite, chaque élément est attribué au cluster dont la centroïde lui est la plus proche et celle-ci est recalculée. Ces deux étapes sont répétées jusqu'à satisfaction d'un critère de convergence (qui est souvent le non changement de l'attribution). Un élément important est que les centroïdes initiales sont sélectionnées de manière aléatoire et, ainsi, la qualité des clusters résultants varie. Les choix initiaux peuvent ainsi conduire à une faible qualité de cluster. Une variante de cet algorithme est le « c-means flou » (*fuzzy c-means* en anglais) (Jain *et al*., 1999), où chaque point est caractérisé par un degré d'appartenance aux clusters (comme en logique floue), au lieu d'appartenir complètement à un seul cluster. Certains points peuvent ainsi appartenir à un cluster à un degré inférieur aux points situés au centre du cluster. Cet algorithme présente le même inconvénient que le *k-means*, dans la mesure où les résultats dépendent du choix initial de poids. L'algorithme d' « espérance-maximisation » (Dempster *et al*., 1977 ; Jain *et al*., 1999 ; Witten et Frank, 2005 : 265) est caractérisé par de meilleures propriétés de convergence que les autres algorithmes, tout en permettant, lui aussi, l'appartenance partielle à des clusters.

Des méthodes hybrides existent également dans la littérature, qui combinent la haute qualité des algorithmes hiérarchiques avec l'efficacité des algorithmes de partitionnement. Tel est, par exemple, l'algorithme utilisé dans l'approche de Schütze (1998), où un algorithme d'agglomération est combiné avec un algorithme d'espérance-maximisation.

1.1.4. Représentation des informations utilisées pour l'apprentissage

1.1.4.1. Espace vectoriel vs *espace de similarité*

Les représentations des objets constituant l'entrée du processus de clustering varient au sein des différentes méthodes. La méthode de « discrimination de groupes de contextes », proposée dans le travail de Schütze (1998), regroupe les instances d'un mot polysémique dans des clusters en fonction de leur similarité contextuelle[71]. Le contexte de ces instances dans le corpus d'apprentissage est représenté à l'aide de **vecteurs**. La méthode opère sur les représentations vectorielles des instances des mots polysémiques, c'est-à-dire dans l'**espace vectoriel** construit[72]. Les vecteurs contextuels constituent l'entrée d'un algorithme de clustering et, ainsi, le regroupement des contextes s'effectue au sein de l'espace vectoriel[73]. Les clusters qui en résultent sont constitués d'instances similaires, d'un point de vue contextuel, et chaque cluster est, par la suite, interprété comme un sens.

Pedersen et Bruce (1997a) représentent les instances dans un **espace de similarité**. Au sein de cet espace, chaque instance est représentée par un **point** et

[71] La méthode de Schütze est une approche de nature « indirecte », dans le sens où elle considère les relations de tous les mots et ne se focalise pas sur un mot cible et ses voisins.

[72] Schütze utilise des vecteurs contextuels de deuxième ordre qui représentent une instance par la moyenne des vecteurs de traits construits pour les mots de contenu qui apparaissent dans le contexte du mot polysémique à cette instance. Nous expliquerons la nature de ces vecteurs dans le paragraphe suivant.

[73] Les clusters sont représentés par leurs centroïdes, c'est-à-dire par la moyenne de leurs éléments. La représentation engendrée est ensuite utilisée pour la désambiguïsation de nouvelles instances des mots polysémiques. Une nouvelle instance est désambiguïsée en calculant la représentation de second degré de son contexte, et en l'attribuant au cluster dont la centroïde est la plus proche de cette représentation.

la distance entre deux points est fonction de leur **similarité**[74]. La matrice qui contient la similarité entre chaque paire d'instances constitue l'entrée d'un algorithme agglomératif de clustering. Un espace de similarité est également construit par la méthode de Pantel et Lin (2002, 2003) et le clustering des mots s'opère au sein de cet espace. La technique utilisée pour le clustering diffère pourtant quelque peu. L'algorithme utilisé forme d'abord un ensemble de clusters forts, appelés « comités », éparpillés au sein de l'espace de similarité. La centroïde de chaque cluster est ensuite construite en trouvant la moyenne des vecteurs de traits d'un sous-ensemble des membres du cluster. Cette centroïde constitue le vecteur de traits du cluster et les mots qui restent sont attribués au cluster dont la centroïde leur est la plus proche[75]. Les clusters finaux correspondent aux différents sens des mots.

Le modèle de Ji *et al*. (2003) organise, lui aussi, les cooccurrents pertinents d'un mot (appelés « contexonymes ») dans un **espace sémantique multi-dimensionnel**. Ce modèle vise la représentation d'informations lexicales de granularité fine et est basé sur le **sens minimal** d'un mot (représenté par une « clique » [76]), ce qui constitue sa différence principale par rapport aux modèles statistiques précédents. L'organisation des contexonymes en cliques reflète l'usage contextuel des mots et leurs liens sémantiques. Elle permet aussi de capter leurs connotations sémantiques de granularité fine et de distinguer entre leurs différents sens. Etant composées de plusieurs ensembles de mots, les cliques sont considérées, dans ce modèle, comme des unités minimales d'un contexonyme, représentant des sens plus fins que le mot lui-même[77].

[74] La similarité entre deux instances est calculée à l'aide du cosinus des vecteurs contextuels qui leur correspondent.

[75] Cette technique ressemble au fonctionnement de l'algorithme « *k*-means », où les éléments sont aussi attribués aux clusters dont les centroïdes leur sont les plus proches. Mais, contrairement à « *k*-means », le nombre des clusters n'est pas fixé à l'avance et les centroides ne sont pas modifiées.

[76] Les cliques sont des sous-graphes complets maximaux.

[77] Le nombre élevé de cliques pour chaque mot rend les différences entre celles-ci trop fines. Par exemple, pour le mot *match*, les auteurs rapportent avoir trouvé 50 contexonymes et 133 cliques. Un contexonyme qui appartient à une seule clique est censé avoir une seule valeur sémantique minimale, tandis que ceux qui appartiennent à un grand nombre de cliques, ont un nombre égal de valeurs sémantiques minimales différentes. C'est probablement pour cette raison que les auteurs soulignent la possibilité de clustering des cliques ou des contexonymes.

1.1.4.2. Graphes de cooccurrence

Les instances décrites au sein d'un espace de similarité, comme dans le travail de Pedersen et Bruce (1997a), peuvent aussi être décrites à l'aide d'un **graphe pondéré**. Dans ce graphe, chaque instance peut être vue comme un nœud d'un graphe pondéré, tandis que le poids de l'arête liant deux nœuds indique leur similarité. Il est assez souvent fait appel dans la littérature à des représentations du contexte lexical sous forme de graphes (Dorow et Widdows, 2003 ; Véronis, 2003, 2004 ; Ferret, 2004a,b; Agirre *et al*., 2006). Les graphes en question sont construits à l'aide des unités lexicales qui se trouvent à proximité des mots étudiés dans les textes et sont ainsi souvent caractérisés comme des **graphes de cooccurrences**. Les mots du corpus correspondent aux nœuds des graphes, tandis que les arêtes représentent les relations de cooccurrence des mots dans les textes[78]. La détection des sens au sein de ces graphes peut être faite par des techniques non supervisées de clustering, comme celles utilisées dans le cas de l'espace vectoriel et de l'espace de similarité. Le clustering des graphes, appelée aussi « partitionnement des graphes », consiste à regrouper des sommets, c'est-à-dire à répartir l'ensemble des nœuds du graphe dans des ensembles disjoints (clusters ou partitions), tout en gardant minimal le nombre d'arêtes liant des nœuds d'ensembles distincts. Les cooccurrences des mots polysémiques trouvées dans le graphe sont clustérisées et les clusters fournis correspondent alors à leurs différents sens.

Dans la méthode de Dorow et Widdows (2003), par exemple, le clustering des mots représentés par les nœuds du graphe se base sur l'observation d'un nombre d'arêtes élevé trouvées à l'intérieur d'une région sémantique, et d'un petit nombre de liens entre régions sémantiques différentes. La détection des régions sémantiques à l'intérieur des graphes locaux, se fait par un algorithme de clustering reposant sur une approche markovienne. Le principe de l'algorithme est que des promenades aléatoires dans le graphe ont plutôt tendance à rester dans le même cluster, et non à aller d'un cluster à l'autre. Il est également supposé qu'un mot polysémique lie des régions sémantiques qui ne seraient pas

[78] Les arêtes liant deux mots peuvent être pondérées en fonction de la fréquence de cooccurrence des mots.

liées autrement. Lorsque ce mot est éliminé de son graphe de cooccurrences, un sous-graphe dont les composantes connexes correspondent aux sens du mot est obtenu. Ainsi, des clusters de sens sont calculés de manière itérative en clusterisant le graphe local de mots similaires autour d'un mot ambigu.

Néanmoins, Véronis (2003 ; 2004) soutient que l'existence de connexions entre les composantes d'un graphe interdit l'utilisation d'algorithmes de détection de composantes fortement connexes ou de cliques. C'est pourquoi, il propose d'isoler des composantes de forte densité à l'intérieur du graphe de cooccurrences, qui correspondent aux différents sens des mots polysémiques. En revanche, Ferret (2004a,b) utilise une adaptation de l'algorithme des plus proches voisins[79] (Ertöz *et al.*, 2001) afin de regrouper les cooccurrents d'un mot polysémique qui définissent un sens (à l'intérieur d'une région de haute densité) dans le sous-graphe correspondant au mot. Au sein de ce sous-graphe, le nombre de relations entre cooccurrents définissant un sens est censé être plus élevé que celui des relations entre cooccurrents définissant des sens différents. Une matrice de similarité des cooccurrents est ensuite construite, en exploitant les relations représentées au sein du sous-graphe.

Les variations observées au niveau de la représentation choisie dans les différentes méthodes d'acquisition de sens ne constituent pas pour autant de différences fondamentales. Les informations incluses dans un graphe peuvent, par exemple, être représentées à l'aide d'un ensemble de vecteurs de traits, construits à partir de la matrice d'adjacence correspondante. Quelle que soit la représentation retenue, l'objectif du clustering reste identique : créer des clusters dont les éléments entretiennent des relations plus fortes entre eux qu'avec les éléments appartenant à d'autres clusters, et qui représentent les différents sens des mots ambigus. Ce but est atteint en utilisant des algorithmes de clustering variés, au sein des différentes méthodes.

[79] Les avantages de cet algorithme pour le clustering sont que le nombre de clusters est déterminé automatiquement et que les éléments non représentatifs des clusters construits ne sont pas pris en considération ; d'où une élimination du bruit.

1.1.5. Résultat des méthodes d'apprentissage

Nous venons d'analyser un ensemble de techniques d'apprentissage utilisées pour le repérage automatique de sens, ainsi que les différentes représentations possibles du contexte lexical pris en compte par ces méthodes. Nous avons montré que les informations contextuelles peuvent être organisées de manières variées, mais que la finalité demeure le clustering d'instances de mots polysémiques dans un corpus de textes, de telle manière que les clusters qui en résultent reflètent les sens véhiculés par ces mots. Les inventaires de sens fournis par ces méthodes peuvent donc constituer une ressource utile pour la désambiguïsation lexicale, c'est-à-dire pour la sélection du sens véhiculé par de nouvelles instances de mots polysémiques. La possibilité de création de tels inventaires à partir de données textuelles constitue précisément un avantage des méthodes automatiques d'acquisition de sens : elles permettent en effet la définition de sens relatifs aux données traitées, ce qui augmente la pertinence des informations sémantiques obtenues et leur utilité lors de la tâche de désambiguïsation.

Cette caractéristique des méthodes automatiques d'acquisition de sens peut être, néanmoins, considérée également comme un de leurs inconvénients. La spécificité des sens obtenus au sujet du corpus dont ils sont dérivés dans un système précis, rend douteuse leur possibilité d'utilisation par d'autres systèmes (Pereira *et al*., 1993). A cela s'ajoute le fort impact du corpus sur la couverture de l'inventaire de sens construit. L'inférence de propriétés en langue à partir d'observations faites sur corpus suppose que le corpus soit à la fois homogène (Habert *et al*., 1997 : 105) et suffisamment vaste pour permettre la description des phénomènes étudiés. La qualité et la fiabilité des descriptions sémantiques augmentent proportionnellement au nombre de contextes dans lesquels figurent les mots, i.e. relativement à la quantité d'information disponible[80]. Les mots faiblement représentés dans les textes sont, par conséquent, les plus difficiles à décrire.

[80] Pour Habert *et al*. (1997: 105), le travail lexicographique ne peut reposer uniquement sur les corpus. Néanmoins, si les informations extraites de corpus sont contrôlées, corrigées et complétées, elles peuvent constituer une vue d'ensemble sur l'emploi d'un mot et une source importante pour la rédaction d'entrées du dictionnaire.

Une autre critique adressée aux méthodes empiriques d'acquisition de sens concerne la difficulté d'interprétation des sens, dans la mesure où il n'est pas toujours évident de savoir à quoi correspondent les sens décrits par les clusters (Pereira *et al*., *ibid*.). A ceci nous pourrions ajouter la granularité très fine des distinctions sémantiques révélées par ces méthodes, surtout lorsqu'elles utilisent des représentations directes du contexte. Ce type de représentations fera l'objet du paragraphe suivant, où nous présenterons une distinction importante, au sein des différentes méthodes, au niveau de la prise en compte des cooccurrences lexicales pour le clustering.

1.2. Prise en compte du contexte lexical

Nous venons de voir comment l'acquisition de sens peut être considérée comme un problème d'apprentissage automatique non supervisé, ainsi que les principes de fonctionnement d'un ensemble de méthodes automatiques développées dans ce but. Ces méthodes se basent sur les régularités distributionnelles découvertes dans les textes pour regrouper les instances des mots ambigus, en fonction de leur similarité. Les informations contextuelles relatives aux mots ambigus peuvent être exploitées de manières variées. Une distinction importante, à ce niveau, concerne la **représentation directe** ou **indirecte** du contexte lexical.

1.2.1. Représentation directe ou indirecte du contexte : cooccurrences d'ordres variés

Une représentation directe du contexte lexical utilise des informations du **contexte immédiat** du mot ambigu au sein d'une fenêtre textuelle, c'est-à-dire ses **cooccurrences directes** (ou cooccurrences de **premier ordre**). Les traits contextuels repérés au sein d'une fenêtre textuelle peuvent être pris en considération d'une manière plus ou moins sophistiquée, sous forme de **sac de mots** ou combinés avec des informations relatives à leur **ordre** et/ou à leur

position dans le texte, ou encore aux **relations** qu'ils entretiennent entre eux[81]. Le repérage des traits contextuels pertinents pour chaque instance du mot ambigu permet la création d'une représentation contextuelle de premier ordre pour chaque instance.

En revanche, une représentation contextuelle indirecte exploite les informations de **cooccurrences indirectes** (ou de **deuxième ordre**), c'est-à-dire les cooccurrences des cooccurrents du mot ambigu[82]. Le principe régissant l'exploitation de ce type de cooccurrences consiste à considérer deux mots ayant des contextes de cooccurrence (de premier ordre) proches comme sémantiquement proches. Dans les deux cas, les représentations contextuelles peuvent être construites soit à partir de l'ensemble des traits du contexte des mots, soit en utilisant un sous-ensemble de ces traits (par ex. les mots de certaines parties du discours).

Les méthodes d'acquisition de sens de Véronis (2003, 2004), de Dorow et Widdows (2003) et de Rapp (2003) reposent sur les cooccurrences directes des mots. Véronis construit des graphes de cooccurrences en ne retenant que les noms et les adjectifs des contextes des mots ambigus ; Dorow et Widdows, eux, n'utilisent que les noms[83], tandis que Rapp élimine simplement les mots fonctionnels des contextes. L'impact de l'utilisation de cooccurrences directes, dans la méthode de Véronis, sur les sens obtenus est évident : les sens obtenus ne correspondent pas à de vrais sens lexicaux mais à différents « usages » des mots en contexte. Cette méthode permet pourtant d'isoler des usages très peu fréquents des mots, ce qui n'est pas le cas des méthodes basées sur des vecteurs. Dans ces méthodes, en effet, une grande différence de fréquence entre usages d'un mot dissimule des distinctions utiles, car situées au-dessous du seuil de bruit adopté par le modèle.

[81] Nous analyserons, dans le paragraphe 1.3. du chapitre 3, les manières dont les informations contextuelles du contexte local peuvent être prises en compte pour la désambiguïsation.

[82] Il est également possible d'avoir des représentations de troisième (quatrième, etc.) ordre, qui exploitent les informations sur les cooccurrences des cooccurrents des cooccurrents (des cooccurrents,…) des mots.

[83] Les contextes considérés sont des listes de noms qui apparaissent dans le corpus, sur la base de l'hypothèse que les noms qui co-apparaissent dans une liste sont généralement sémantiquement liés.

Cette possibilité de mise en évidence de sens peu fréquents des mots, qu'offrent les associations de premier ordre, est soulignée aussi par Rapp (2003)[84]. Dans cette méthode, les associations de premier ordre les plus fortes du mot ambigu sont repérées en utilisant le logarithme de la vraisemblance et une fenêtre de petite taille autour de celui-ci (± 2 mots). Un score de « complémentarité » [85] est calculé pour chacune des paires possibles de ces associations et les mots de la paire qui obtient le score le plus élevé sont alors considérés comme étant les meilleurs descripteurs des deux principaux sens du mot ambigu[86].

1.2.2. Vecteurs d'ordres variés

Dans une grande partie des travaux d'acquisition de sens, le contexte des mots ambigus est représenté à l'aide de vecteurs. Tout comme une distinction est établie entre cooccurrences de « premier » et de « deuxième ordre », il est possible de distinguer deux types de vecteurs contextuels en fonction des traits utilisés pour leur construction : des vecteurs contextuels de « premier » ou de « deuxième ordre ». Les **vecteurs de premier ordre** représentent le contexte de chaque instance des mots ambigus en utilisant, de manière directe, les traits qui le constituent (c'est-à-dire ses cooccurrences directes). Purandare et Pedersen (2004a) utilisent ce type de vecteurs, construits à partir d'un petit nombre de traits locaux, qui concernent la cooccurrence des mots à l'intérieur d'une fenêtre textuelle de vingt mots de contenu autour du mot ambigu. Des cooccurrences de premier ordre sont aussi utilisées par Pantel et Lin (2002, 2003), mais la nature des traits contextuels retenus est différente. Dans cette méthode, chaque mot est représenté par un vecteur de traits, où chaque trait correspond à un contexte

[84] Selon Rapp, les associations de premier ordre reflètent, le plus souvent, la totalité des sens, contrairement aux associations de « deuxième ordre » – que nous décrirons par la suite – qui reflètent uniquement le sens principal du mot.

[85] Pour que les mots d'une paire soient de bons descripteurs de sens, ils doivent être les plus dissemblables possibles et la somme de leurs vecteurs de cooccurrence doit correspondre au vecteur du mot ambigu. Ainsi, la préférence est attribuée aux paires dont les cooccurrents sont complémentaires.

[86] La méthode de Rapp (2003) vise à distinguer entre les deux sens principaux des mots ambigus contenus dans la liste utilisée par Yarowsky (1995). Yarowsky fournit un nom caractéristique pour chaque sens de ces mots, c'est-à-dire un descripteur. L'évaluation de la méthode de Rapp est faite en comparant les descripteurs de sens repérés avec ceux donnés par Yarowsky.

syntaxique dans lequel le mot apparaît (par exemple, un contexte 'verbe-objet'). Chaque trait est pondéré par l'information mutuelle ponctuelle calculée entre le trait et le mot[87].

Les **vecteurs de deuxième ordre** donnent la possibilité de représenter le contexte de manière indirecte, en exploitant les informations de cooccurrence de deuxième ordre. Ces vecteurs sont construits sur la base d'une moyenne des vecteurs de premier ordre correspondant aux traits contextuels des mots. Des vecteurs de ce type sont utilisés par les méthodes de Schütze (1998) et de Ferret (2004a). Dans la méthode de Schütze, chaque trait du contexte des mots est représenté, initialement, comme un vecteur de ses propres cooccurrents dans les données d'apprentissage. Ces vecteurs sont les vecteurs contextuels de premier ordre des mots constituant les traits (et non du mot polysémique). Les cooccurrents des traits forment les dimensions d'un espace vectoriel ; un espace de grandes dimensions est ainsi formé[88]. Une instance d'un mot est représentée par un vecteur contextuel de deuxième ordre, qui correspond à la moyenne des vecteurs de premier ordre associés à ses cooccurrents (traits contextuels du mot).

1.2.3. Avantages et inconvénients d'une représentation directe ou indirecte du contexte

Des expériences menées par Purandare et Pedersen (2004a,b) ont montré que les vecteurs de premier ordre sont plus sensibles à l'effet de la **dispersion des données** (*data sparseness*). Le problème devient même plus aigu avec des corpus d'apprentissage de petite taille, où la dispersion des données est grande et l'ensemble de traits contextuels repérés petit. En effet, dans le cas de petites quantités de données, il n'est guère probable de trouver des instances des mots, dans le contexte desquelles figurent les mêmes mots. Des instances sémantiquement liées apparaissent dans des contextes où des mots similaires sont utilisés mais cette similarité est conceptuelle et non lexicale. Le faible taux de

[87] La similarité entre deux mots est calculée à l'aide du cosinus des vecteurs correspondants. Les mots les plus similaires sont d'abord clusterisés et ensuite, chaque mot est attribué aux clusters qui lui sont le plus similaires. Les clusters auxquels un mot est attribué décrit ses différents sens.

[88] Les traits sont sélectionnés, dans ce travail, sur la base de leur fréquence ou du logarithme du rapport de vraisemblance au sein du corpus d'apprentissage.

correspondance entre les contextes des instances des mots ambigus rend le clustering difficile[89]. Les vecteurs de deuxième ordre sont désignés pour identifier des relations de similarité conceptuelle. Dans ce cas, une correspondance exacte entre les mots des contextes n'est pas requise, mais les mots apparaissant dans des contextes similaires sont censés avoir des vecteurs similaires.

Les vecteurs de deuxième ordre représentant les traits du contexte de manière indirecte, les méthodes qui utilisent ce type de vecteurs fonctionnent donc mieux sur de petits ensembles de données, car les relations de cooccurrence de deuxième ordre fournissent des informations suffisantes pour la discrimination des sens. En revanche, lorsque l'ensemble des données d'apprentissage est important, un grand espace de traits est construit, ce qui augmente les possibilités de correspondance directe avec les traits trouvés dans les contextes des instances des mots ambigus. Dans ce cas, les vecteurs de premier ordre sont alors plus appropriés puisqu'il est davantage probable de trouver des correspondances exactes entre les mots utilisés au sein des contextes.

Ferret (2004a) souligne également que l'utilisation des cooccurrences de deuxième ordre permet de capter des relations sémantiques difficiles à capter à l'aide des seules cooccurrences textuelles. Dans ce travail, un vecteur de taille égale au nombre de cooccurrents du mot considéré est associé à chacun des cooccurrents. Ce vecteur contient les valeurs de cohésion liant les cooccurrents entre eux, calculées par l'information mutuelle. Une matrice de similarité est construite en calculant le cosinus de chaque paire de vecteurs, correspondant à chaque paire de cooccurrents. Par cette mesure, la similarité entre deux cooccurrents peut être définie, même si ces cooccurrents ne sont pas directement liés au sein du réseau de cooccurrences, mais seulement par le nombre significatif de mots qui leur sont communs.

Nous venons de montrer que les méthodes d'acquisition de sens opératoires dans un cadre monolingue se basent sur des régularités distributionnelles pour repérer les sens lexicaux. Après avoir présenté les principales techniques d'apprentissage utilisées pour l'acquisition de sens à partir de textes et certaines

[89] Dans le cas de corpus de petite taille, il serait donc souhaitable, d'après Purandare et Pedersen, d'enrichir l'ensemble de traits avec des données provenant d'autres sources d'apprentissage.

manières de considérer le contexte lexical au sein des différents modèles, nous allons désormais, dans le paragraphe suivant, étudier un ensemble de méthodes d'acquisition de sens proposées dans un cadre bi- (ou multi-)lingue, i.e. impliquant l'utilisation de plus d'une langue.

2. Repérage automatique de sens dans un cadre bi- (multi-) lingue

2.1. Méthodes traductionnelles d'acquisition de sens

2.1.1. Principes sous-jacents aux méthodes traductionnelles

Les hypothèses sous-jacentes au fonctionnement des méthodes d'acquisition de sens développées dans un cadre bilingue ou multilingue diffèrent de manière importante de celles sur lesquelles se basent les méthodes monolingues. En effet, les méthodes bi- ou multilingues exploitent généralement des informations de traduction ; d'où leur dénomination comme **méthodes traductionnelles**. Le principe fondamental gouvernant ces méthodes est que les différents équivalents de traduction d'un mot polysémique de la langue source constituent des indices des distinctions sémantiques du mot. Par conséquent, les équivalents peuvent être utilisés pour repérer les sens du mot polysémique, en en révélant les sens « cachés » et, éventuellement, pour étiqueter les instances de ce mot à l'aide des sens rendus évidents par l'analyse sémantique.

Des exemples courants dans la littérature concernent les mots anglais *duty* et *bank*, dont les sens principaux sont révélés à l'aide de leurs équivalents de traduction en français (respectivement *droit-devoir* et *banque-rive*) (Resnik et Yarowsky, 2000 ; Resnik, 2004). Des conclusions sémantiques peuvent également être déduites dans le cas de correspondance entre différents mots source et un seul équivalent de traduction. Ce type de correspondance traductionnelle indique, souvent, que les mots source partagent un élément de sens (Resnik, 2004)[90]. Les traductions des mots sont ainsi considérées comme des « fenêtres » sur leurs propriétés sémantiques et constituent une base empirique pour leur description.

[90] Par exemple, dans le cas des mots anglais *bank* et *shore*, le fait qu'ils puissent être tous les deux traduits par *rive* en français suggère que les deux sens qui correspondent à cette traduction partagent une même propriété sémantique.

2.1.2. Avantages des méthodes traductionnelles

2.1.2.1. Les traductions: une source objective d'informations sémantiques

L'utilisation des traductions en tant que source pour l'identification de distinctions sémantiques est considérée comme étant une solution au problème de la **subjectivité** qui caractérise la définition des sens dans les ressources existantes, subjectivité qui apparaît dans les divergences observées entre ces ressources[91]. Les variations concernant la conception du sens et la finalité des ressources constituent des facteurs qui contribuent au manque d'uniformité tant au niveau des descriptions sémantiques entre ressources différentes, qu'au niveau des jeux d'étiquettes utilisés lors des tâches d'étiquetage sémantique.

Les traductions sont, en revanche, considérées comme étant une **source objective** d'informations sémantiques (Resnik et Yarowsky, 2000 ; Ng *et al.*, 2003). La relation de traduction peut même être considérée comme un **primitif théorique** (Dyvik, 1998a, 2003, 2005), i.e. un concept qui n'est pas défini en termes d'autres concepts, mais qui peut être extrait de données traductionnelles par des méthodes interprétatives. Les données de traduction issues d'un corpus parallèle peuvent alors être regardées comme le résultat d'un processus au cours duquel le mot source a été interprété dans son contexte[92]. Par conséquent, les relations sémantiques résultant de traductions sont considérées comme des relations ne découlant pas de considérations philosophiques et/ou théoriques sur le sens[93]. Ce souci d'objectivité au niveau des descriptions sémantiques est d'ailleurs visible dans des travaux comme celui de Resnik et Yarowsky (*ibid.*), qui proposent de restreindre un inventaire sémantique d'une LS aux distinctions

[91] Nous avons déjà abordé la problématique de la définition des sens lexicaux, de leur nombre et de leur granularité, et nous avons décrit un ensemble de contraintes extra-linguistiques qui interviennent lors de la construction d'une ressource (ainsi, l'objectif visé et les utilisateurs envisagés, ou l'application visée dans un cadre automatique).

[92] Les traducteurs évaluent les possibilités interprétatives des expressions linguistiques de la LS dans des contextes spécifiques, des textes ayant des objectifs précis, et essaient, ensuite, de récréer les mêmes possibilités d'interprétation dans un texte cible, qui sert un objectif comparable dans une autre langue.

[93] Les études sémantiques dépendent souvent de paraphrases, ou de manières alternatives de dire la même chose. Les traductions constituent une source de telles alternatives, théoriquement intacte, dans le sens où celles-ci peuvent être extraites de corpus et constituent ainsi des données empiriques.

sémantiques lexicalisées dans d'autres langues[94]. Kaji (2003), tout en prenant en considération la non univocité de correspondance entre les sens d'un mot polysémique et ses équivalents de traduction, propose, quant à lui, l'élaboration d'un inventaire sémantique où chaque sens serait défini comme un ensemble d'équivalents de traduction synonymes dans une autre langue[95].

2.1.2.2. Création automatique de corpus sémantiquement étiquetés

Un autre avantage des méthodes traductionnelles de repérage de sens et de désambiguïsation est d'offrir la possibilité de création automatique de **corpus sémantiquement étiquetés**, où les mots sont étiquetés à l'aide des équivalents de traduction qui servent à repérer leurs sens. L'intérêt de ces méthodes est grand dans la mesure où créer manuellement des ressources de ce type requiert beaucoup de temps et met en jeu la question de la subjectivité des annotateurs. Des expériences menées sur le sujet ont montré un fort taux de désaccord entre annotateurs, dû, en grande partie, à la nature des distinctions sémantiques fournies dans les ressources dont ils se servent pour effectuer l'étiquetage (Ng *et al.*, 1999 ; Ide *et al.*, 2001). Cette difficulté, constatée pour l'acquisition de ressources étiquetées de grandes dimensions et de bonne qualité, ne diminue pas pour autant leur grande utilité dans certaines applications du TAL, comme la désambiguïsation lexicale supervisée où elles pourraient constituer la base de l'apprentissage (Ide *et al.*, 2001 ; Diab et Resnik, 2002 ; Ng *et al.*, 2003 ; Resnik, 2004 ; Lyse, 2006).

Ce type de processus d'étiquetage non supervisé, fondé sur les traductions, présente l'avantage de ne pas nécessiter d'inventaire sémantique prédéfini dans la langue du corpus à étiqueter pour l'apprentissage. Néanmoins, certaines de ces

[94] Cette solution se situerait à mi-chemin entre les distinctions grossières – comme celles trouvées au niveau des homographes – et l'expression de distinctions de granularité très fine. D'un point de vue pratique, Resnik et Yarowsky proposent la définition d'un ensemble de langues (les expériences menées par eux impliquent 12 langues différentes) et l'utilisation de dictionnaires bilingues associés pour le repérage des traductions. Ainsi, chaque distinction sémantique doit être réalisée lexicalement dans un sous-ensemble minimal des langues choisies. L'utilisation d'un vaste ensemble de langues est considérée comme ayant un effet positif sur la qualité des résultats de l'acquisition des sens.

[95] Cet élément différencie la méthode d'acquisition de sens proposée par Kaji des méthodes monolingues d'acquisition de sens, où les sens sont souvent définis comme un ensemble de synonymes au sein de la même langue.

méthodes n'utilisent pas directement les équivalents comme étiquettes des mots source, mais ont plutôt recours à un inventaire dans la LC. L'algorithme de Diab et Resnik (2002), par exemple, regroupe les traductions d'un mot source, récupère leurs étiquettes sémantiques possibles à partir d'un inventaire, puis sélectionne le sens qui caractérise l'ensemble de ses traductions (ou leurs sens les plus proches) et qui sert à étiqueter le mot source. Cette méthode se fonde sur l'hypothèse de **monosémie** des mots source, manifeste dans le principe de repérage d'un sens commun à ses différentes traductions. Les cas de **mots sémantiquement distants** alignés au même mot source ne sont donc pas pris en considération, bien que les auteurs soulignent la fréquence de ce phénomène. L'amélioration envisagée à propos du fonctionnement de cet algorithme consiste à incorporer des informations de cooccurrence pour **clustériser** les traductions, dans le but de distinguer les sens des mots source[96].

2.1.2.3. Conformité pour le traitement bi- (et multi-) lingue

L'utilisation de méthodes traductionnelles pour l'analyse sémantique présente également des avantages au niveau des applications. Une critique régulièrement émise à l'égard des ressources sémantiques préétablies est qu'elles ne répondent pas aux besoins d'applications réelles. L'établissement de distinctions sémantiques par le biais des traductions différentes des mots, dans le cas des méthodes traductionnelles d'analyse sémantique, rend ces dernières conformes aux besoins de traitement dans le cadre d'applications bilingues ou multilingues. Par exemple, les distinctions sémantiques repérées peuvent être utilisées dans des tâches de désambiguïsation lexicale. La nature de ces distinctions permet en effet la sélection automatique d'un mot de la LC à la sortie de l'étape de désambiguïsation d'un mot source, ce qui correspond précisément à la sélection lexicale effectuée dans la Traduction Automatique (Ng *et al.*, 2003)[97].

[96] Diab et Finsh (2000) utilisent également des techniques de clustering pour la création de correspondances au niveau des mots dans des corpus comparables.

[97] Nous reviendrons sur l'assimilation de la tâche de désambiguïsation et de la tâche de sélection lexicale en §1.2. du chapitre 8, qui porte sur le besoin de désambiguïsation lexicale pour la traduction automatique.

Nous allons désormais analyser certains facteurs qui conditionnent le bon fonctionnement des méthodes traductionnelles d'acquisition de sens. Mais auparavant, nous estimons nécessaire de clarifier la manière dont la notion de « contexte lexical » est conçue dans un cadre bilingue et multilingue, qui diffère de celle définie dans un cadre monolingue. Ainsi, cette clarification permettra d'éviter des confusions et de mieux comprendre le fonctionnement des méthodes développées dans ces deux cadres.

2.2. Le « contexte lexical » bi- (multi-)lingue

2.2.1. Conception de la notion de contexte dans un cadre de traduction

Rappelons que les méthodes contextuelles monolingues d'acquisition de sens, présentées dans le paragraphe 1, exploitent les informations de cooccurrence venant du contexte local des mots. Ces informations peuvent être plus ou moins sophistiquées et concernent les mots qui co-apparaissent au sein d'une fenêtre textuelle, plus ou moins grande, autour des mots ambigus, ou qui entrent dans certains types de relations avec eux.

Dans un cadre impliquant l'utilisation de plus d'une langue, la notion de « contexte lexical » peut être conçue autrement. Le plus souvent, dans un tel cadre, le contexte des mots ne correspond pas à leur contexte lexical à l'intérieur de la même langue mais à leurs **traductions** dans d'autres langues, au sein de corpus parallèles, ou correspond au contexte de ces traductions. Cette conception du contexte se retrouve dans de nombreuses méthodes traductionnelles d'analyse sémantique, comme celles d'Ide *et al*. (2001, 2002), Tufiş *et al*. (2004c), Kaji (2003) et van der Plas et Tiedemann (2006). Elle repose sur l'hypothèse de **lexicalisation** différente des sens d'un mot dans d'autres langues[98]. Sur la base de cette hypothèse, la traduction est supposée « capter », d'une certaine manière, le contexte de la LS tel que le traducteur l'a conçu et, éventuellement, tel qu'il l'a utilisé pour identifier le sens correct du mot source.

[98] Hypothèse qui n'est vraie que jusqu'à un certain point, en raison de l'éventuelle préservation de l'ambiguïté entre les langues. La préservation de l'ambiguïté dépend d'un ensemble de paramètres, comme la typologie des langues et la distance entre elles (Ide, 1999 ; Ide *et al*., 2002 ; Tufiş *et al*., 2004).

Les informations contextuelles de ce type peuvent être utilisées d'une manière similaire à celle employée dans les méthodes monolingues, pour le clustering sémantique des instances des mots et l'identification de distinctions sémantiques. Dans certaines méthodes, elles constituent même la seule source d'informations pour l'analyse sémantique, tandis que dans d'autres, elles sont enrichies par des informations provenant d'autres ressources. Tel est le cas, par exemple, dans la méthode de Tufiş *et al*. (*ibid*.), où les informations de traduction sont complétées par des informations du réseau sémantique multilingue BalkaNet[99]. Nous allons montrer maintenant la manière dont le clustering des instances et le repérage des sens sont réalisés au sein des méthodes traductionnelles précitées.

2.2.2. Clustering au sein de méthodes traductionnelles

La méthode d'Ide *et al*. (2001, 2002) considère comme contexte lexical d'un mot, ses traductions dans un grand ensemble de langues. Chaque mot polysémique source (anglais) est associé à l'ensemble de ses traductions au sein d'un **corpus parallèle** aligné, composé de versions du même texte dans six langues différentes (roumain, slovène, tchèque, bulgare, estonien et hongrois). Le contexte lexical du mot ambigu correspond à ses traductions trouvées au sein du corpus parallèle. Un **vecteur** est construit pour chaque occurrence du mot dans le corpus, qui représente les traductions de cette occurrence précise dans les six langues[100]. Les vecteurs créés constituent l'entrée d'un algorithme d'agglomération, qui les clusterise sur la base de la distance minimale calculée entre eux et fusionne, de manière itérative, les paires de clusters. Les clusters finaux représentent les différents sens et sous-sens du mot ambigu source, à l'instar du clustering appliqué dans un cadre monolingue (Schütze, 1992 ; 1998)[101].

[99] Nous décrirons cette méthode en détail dans le paragraphe 2.2.2. du chapitre 3, qui porte sur la désambiguïsation lexicale dans un cadre bi- et multi-lingue.

[100] Si un équivalent donné est utilisé pour traduire une occurrence *i* du mot polysémique dans le corpus, le vecteur a 1 en position *i*, sinon 0.

[101] Les clusters dérivés de cette manière n'identifient que les instances plus ou moins proches, sans fournir une description du sens comme celle qui serait fournie par un dictionnaire, ni choisir des étiquettes sémantiques d'une liste prédéfinie. Pourtant, d'après Ide *et al*. (2001), ceci ne constitue

Le travail de van der Plas et Tiedemann (2006) se base également sur l'alignement multilingue des mots. Les **contextes d'alignement** dans lesquels un mot est trouvé au sein du corpus utilisé, et qui correspondent aux mots des autres langues avec lesquels il est aligné, constituent les traits du vecteur correspondant à ce mot. Le vecteur construit est appelé **vecteur contextuel**, de la même manière que dans les travaux où les vecteurs sont construits à partir des informations de cooccurrence. Les vecteurs d'alignement sont comparés entre eux et leur similarité montre la similarité distributionnelle des mots. Comme dans un cadre monolingue, les mots qui présentent une similarité distributionnelle, qui partagent donc un certain nombre de contextes traductionnels, sont considérés comme sémantiquement liés.

Les travaux de Kaji et Morimoto (2002) et Kaji (2003) combinent, quant à eux, informations contextuelles monolingues et informations traductionnelles pour l'acquisition de sens. La conception du contexte adoptée dans cette approche ressemble cependant davantage à celle des méthodes monolingues. Cette différence avec les autres méthodes multilingues de désambiguïsation, décrites ci-dessus, s'explique par le fait que les corpus utilisés sont des **corpus comparables**, c'est-à-dire des corpus monolingues qui n'entretiennent pas de relations de traduction[102]. Plus précisément, la similitude de cette approche avec l'approche contextuelle « classique » consiste dans le repérage de relations entre les mots au sein de chaque langue, à l'aide d'une mesure d'association des mots (l'information mutuelle). Des paires de mots liés sont donc, tout d'abord, extraites de chaque langue à l'aide d'informations contextuelles monolingues qui sont, par la suite, mises en correspondance sur la base d'informations de traduction issues d'un dictionnaire. Ensuite, pour chaque paire de mots liés d'une langue, un ensemble de paires correspondantes est défini dans l'autre langue représentée dans le corpus, et pour chaque alignement de paires de mots,

pas une faiblesse de la méthode ; la désambiguïsation n'a pas besoin de faire appel à ce type de connaissances (définitions de sens), les informations sur l'utilisation d'un ensemble d'instances d'un mot ambigu dans le même sens (ou dans un sens différent) étant souvent suffisantes.

[102] L'exploitation de corpus comparables par la méthode de Kaji constitue une de ses forces, étant donné que la disponibilité de tels corpus est beaucoup plus importante que celle de corpus parallèles. Néanmoins, cet avantage est nuancé par l'inconvénient de devoir utiliser un dictionnaire bilingue pour l'extraction des relations de traduction, dans la mesure où l'alignement lexical est très difficile dans le cas de corpus comparables. Ainsi, la méthode est soumise aux limitations inhérentes à l'exploitation de ressources lexicales préétablies.

un ensemble de mots liés communs est construit[103]. L'hypothèse sous-jacente à la méthode présuppose que les traductions de mots liés dans une langue correspondent à des mots également liés dans l'autre langue (Rapp, 1995).

Dans cette méthode, les sens d'un mot sont initialement définis à l'aide des équivalents de traduction différents, puis les sens sont progressivement agglomérés sur la base de motifs distributionnels qui montrent leur similarité. Les sens d'un mot sont décrits par des **ensembles de synonymes**, constitués du mot lui-même et d'un ou plusieurs équivalents de traduction représentant le sens en question dans une autre langue. Les équivalents de traduction synonymes d'un mot sont supposés posséder des motifs de distribution similaires, ce qui permet leur **clustering**[104]. La corrélation entre un sens d'un mot polysémique et un indice contextuel est calculée sur la base, d'une part, de leur information mutuelle et d'autre part, de la plausibilité des alignements (entre la paire mot-indice de la LS et une paire correspondante dans la LC) proposant le sens. Cette plausibilité est définie comme la somme pondérée des corrélations entre le sens et les mots liés communs. Cette méthode pourrait être caractérisée comme une **méthode distributionnelle interlangue de clustering** : un mot n'est pas caractérisé par un vecteur construit à partir des mots de la même langue, comme c'est le cas dans le clustering distributionnel conventionnel, mais par un ensemble pondéré de mots de l'autre langue.

Ce type de clustering, effectué dans un cadre bilingue, ne doit pas être confondu avec le clustering utilisé afin d'améliorer la qualité de l'alignement lexical. Dans ce dernier type de travaux, la formation de classes (clusters) de mots fournit une solution au problème de la dispersion des données. L'utilisation de classes permet d'effectuer des généralisations à partir des données et d'éliminer le besoin de correspondances exactes entre données d'apprentissage et données

[103] Pour qu'un mot fasse partie de cet ensemble de mots liés communs aux paires des deux langues, il faut d'abord qu'il soit statistiquement lié aux deux mots de la paire de la LS. Il faut, ensuite, qu'une correspondance existe dans le dictionnaire entre ce mot et un mot de la LC et également que ce dernier soit statistiquement lié aux mots de la paire correspondante de la LC.

[104] L'algorithme de clustering utilisé ne permet pas le chevauchement des clusters. Le choix de ce type d'algorithme repose sur l'hypothèse qu'un équivalent de traduction ne représente souvent qu'un sens du mot ambigu, au moins dans le cas où les deux langues ont des origines différentes (les expériences menées par Kaji concerne la paire de langues anglais-japonais). La méthode de clustering ne permet pas non plus à un indice d'être lié à plus d'un sens, suivant en cela l'hypothèse d'« un sens par discours » (Gale *et al.*, 1992), selon laquelle un mot est toujours employé dans le même sens au sein d'un document.

d'entrée du système, afin d'aboutir à une solution. La formation de classes de mots des deux langues, liées entre elles, permet d'**aligner des ensembles de mots**, au lieu d'aligner des mots, processus supposé diminuer l'impact de la **dispersion des données** sur le résultat de la TA (Och et Weber, 1998 ; Och, 1999).

Le clustering des mots effectué au sein de ces travaux se base sur l'algorithme d'espérance-maximisation. Deux types de probabilité sont modélisés : premièrement, une probabilité monolingue *a priori*, qui concerne l'appartenance d'un mot source à une classe et une probabilité de bigrammes, c'est-à-dire de transition d'une classe à une autre ; deuxièmement, une probabilité de traduction des mots d'une classe de la LS par les mots d'une classe de la LC, sur la base des résultats de l'alignement lexical. Les classes de mots liées des deux langues, ainsi obtenues, sont ensuite exploitées dans le but de généraliser l'applicabilité des patrons d'alignement, utilisés pour trouver la meilleure traduction possible d'une nouvelle phrase de la LS. Le clustering des mots des deux langues sert donc un but bien précis et n'implique pas de considérations d'ordre sémantique.

Nous venons de présenter la manière dont la notion de contexte lexical est prise en compte par les méthodes d'acquisition de sens développées dans un cadre bi-(ou multi-) lingue. L'inventaire de sens fourni par ces méthodes constitue bien souvent une ressource utile pour la désambiguïsation lexicale, i.e. pour la sélection du sens véhiculé par de nouvelles instances des mots polysémiques. La possibilité d'élaborer des inventaires sémantiques à partir des données présente l'avantage de la définition de sens relatifs aux données traitées, ce qui augmente la pertinence des informations sémantiques obtenues. En outre, comme nous l'avons déjà souligné, les méthodes de repérage de sens opératoires dans un cadre multilingue fournissent des résultats directement utilisables dans des applications relatives au traitement multilingue, à la désambiguïsation lexicale et également, à la sélection lexicale dans le cadre d'applications de traduction.

La performance des méthodes traductionnelles d'acquisition de sens, décrites dans le §2.1, dépend néanmoins de facteurs relatifs aux langues utilisées, à leur distance typologique et à leurs relations. Dans le paragraphe suivant, nous

allons étudier plus précisément ces facteurs et analyser leur impact sur le processus de repérage de sens.

2.3. Paramètres conditionnant la réussite des méthodes traductionnelles d'acquisition de sens

2.3.1. Ambiguïté traductionnelle

L'hypothèse principale sur laquelle se basent les méthodes traductionnelles d'acquisition de sens considère que les équivalents de traduction des mots polysémiques peuvent servir d'indices au repérage de distinctions sémantiques au sein de ces mots. Le bon fonctionnement de ces méthodes dépend donc de manière importante du taux de **relations biunivoques** existant entre les équivalents et les sens du mot polysémique, c'est-à-dire de la possibilité de mise en correspondance de chaque équivalent à un seul sens du mot source. Cependant, le repérage de telles relations n'est pas toujours évident et ce, en raison de l'existence possible de relations d'**ambiguïté parallèle** (ou **ambiguïté traductionnelle**) entre les mots de deux langues. Les phénomènes de distinction complète des sens et d'ambiguïté traductionnelle ont déjà été analysés dans le §2.3. du chapitre 1, qui décrit les correspondances de traduction possibles entre les mots polysémiques d'une langue et leurs équivalents de traduction dans une autre.

Les correspondances sémantiques possibles entre mots polysémiques de deux langues peuvent être très complexes. L'établissement de la totalité des correspondances sémantiques possibles entre les mots de deux langues nécessiterait une analyse complète de la sémantique des équivalents de traduction, ce qui peut être évité lors de l'acquisition de sens. En se focalisant sur les mots polysémiques d'une seule langue (le plus souvent la LS), la sémantique des équivalents peut en effet être « restreinte » aux sens liés à ceux des mots source[105].

[105] Comme nous l'avons déjà dit, une telle restriction est facile à réaliser dans une étude basée sur corpus, en considérant uniquement les instances des équivalents qui traduisent des instances du

En revanche, les relations de **recouvrement sémantique** partiel ou total, entre les mots de deux langues peuvent poser problème lors d'un processus d'acquisition de sens reposant sur les traductions. Ces relations empêchent en effet l'utilisation des équivalents comme indices pour l'analyse sémantique des mots source. La solution, dans ces cas, consiste à recourir à des traductions dans d'autres langues, afin d'y découvrir des lexicalisations des sens des mots source. Ainsi, Ide *et al*. (2001, 2002) se fondent sur les traductions d'un mot source (anglais) dans six langues différentes (roumain, slovène, tchèque, bulgare, estonien et hongrois), et Resnik et Yarowsky (2000) mènent des expériences impliquant douze langues différentes (basque, japonais, coréen, chinois, turc, hongrois, roumain, grec, hindi, arabe, espagnol, suédois). Les distinctions sémantiques retenues sont celles qui sont réalisées lexicalement dans un sous-ensemble minimal des langues choisies. L'utilisation d'un nombre élevé de langues est considérée comme ayant un effet positif sur la qualité des résultats de l'acquisition de sens, parce qu'elle permet justement de surmonter l' « obstacle » des ambiguïtés traductionnelles qui peuvent exister entre les mots de deux langues.

Un autre paramètre étroitement lié à la fréquence du phénomène de l'ambiguïté traductionnelle est la distance entre les langues étudiées. C'est ce paramètre que nous allons désormais analyser.

2.3.2. Le paramètre de la distance inter-langue

La **distance historique** et **typologique** entre langues conditionne l'existence de relations d'ambiguïté parallèle entre les mots et donc, la possibilité d'utilisation des équivalents de traduction des mots polysémiques en tant qu'indices pour leur analyse sémantique. Cette distance inter-langue constitue ainsi un facteur important de réussite des méthodes traductionnelles d'acquisition de sens.

Une réponse intuitive à la question de l'impact de ce facteur sur le repérage des sens consisterait à dire que la **lexicalisation** des sens est plus fréquente entre

mot source. De cette manière, lorsque les équivalents expriment des sens non exprimés par l'unité source, une part de leur polysémie peut être ignorée.

langues distantes, tandis que les mots de langues proches présentent davantage de recouvrement au niveau des sens. De nombreux travaux (Ide, 1999a, 1999b ; Resnik et Yarowsky, 2000 ; Dyvik, 2003 ; Ide *et al*., 2002 ; Tufiş et Ion, 2003) explorent la validité d'une telle hypothèse. Il est intéressant de noter que ces expériences fournissent des résultats relativement variés.

Afin de mesurer l'influence de la distance inter-langue sur la lexicalisation des distinctions sémantiques, Resnik et Yarowsky (2000) ont mené des expériences sur l'anglais et douze langues différentes, réparties dans deux groupes : les langues indo-européennes (roumain, grec, hindi, espagnol, suédois) et les langues non indo-européennes (basque, japonais, coréen, chinois, turc, hongrois, arabe). Confirmant l'hypothèse, les résultats obtenus montrent que les langues du deuxième groupe, les plus distantes de l'anglais, présentent une plus grande tendance à lexicaliser les distinctions sémantiques observées en anglais que les langues indo-européennes. En outre, il a été démontré que la distance inter-langue est également liée à la **granularité** des distinctions sémantiques lexicalisées. A un niveau de granularité grossier, les distinctions repérées au niveau des homographes, au sein de la LS, se reflètent sur des unités lexicales différentes tant dans les langues proches que dans les langues distantes. Effectivement, la préservation de l'ambiguïté au niveau des homographes d'une langue à l'autre et l'existence d'ambiguïté traductionnelle sont rares[106].

En revanche, au niveau des distinctions sémantiques plus fines (sens principaux d'un polysème, sous-sens, sous-sous-sens, etc.), les langues indo-européennes ont tendance à présenter des ambiguïtés parallèles à l'anglais et lexicalisent ces distinctions avec une probabilité plus petite que les langues distantes. Les résultats de cette expérience montrent, par conséquent, que l'utilisation de langues distantes donnerait de meilleurs résultats dans un cadre de désambiguïsation et d'annotation sémantique à tous les niveaux, y compris celui des distinctions sémantiques fines.

Les résultats des expériences menées par Ide (1999a, 1999b) diffèrent considérablement de ceux présentés par Resnik et Yarowsky (2000). Les expériences d'Ide portent sur cinq langues de quatre familles

[106] Pour Dyvik (2003, 2005), ceci s'explique par le fait que l'ambiguïté contrastive est une propriété historiquement accidentelle et idiosyncrasique des mots, dont il ne faut pas s'attendre à trouver des instances dans d'autres langues.

différentes (l'anglais : langue germanique ; le tchèque et le slovène : langues slaves ; l'estonien : langue fino-ougrienne (non indo-européenne) et le roumain : langue romane)[107]. L'objectif de ces expériences est, plus concrètement, de définir l'importance à attribuer aux différences de lexicalisation existant entre langues proches (au sein desquelles les ambiguïtés sont généralement préservées), et langues distantes. L'objectif est par ailleurs de définir le nombre de langues nécessaires pour fournir des informations suffisantes à l'analyse, ainsi que les langues en question[108].

De fait, la différence entre les résultats obtenus par Ide et ceux obtenus par Resnik et Yarowsky (*ibid*.) est frappante. Ide soutient que la tendance de lexicalisation des sens n'est pas affectée par la distance inter-langues, contrairement aux conclusions tirées par Resnik et Yarowsky, d'après lesquelles, nous rappelons, les langues non indo-européennes présentent une tendance plus forte de lexicalisation des distinctions sémantiques présentes en anglais que les langues indo-européennes, et ce, à tous les niveaux de granularité.

Cette divergence entre résultats peut néanmoins être attribuée, selon Ide, à la nature des données utilisées dans les deux expériences. Les données de traduction utilisées dans l'étude de Resnik et Yarowsky étaient créées par des locuteurs natifs, à qui il a été demandé d'étiqueter les instances de mots polysémiques, incluses dans des phrases isolées en anglais, par une seule traduction dans leur langue maternelle. Pour une partie des exemples – qui consistait en des paires de phrases contenant des instances véhiculant deux sens différents du mot (de granularité variable) – il a été explicitement demandé aux annotateurs d'identifier s'il y avait une paire de mots dans leur langue qui distinguait les deux sens (des traductions non interchangeables pour les deux sens). Ce processus a ainsi directement favorisé les distinctions lexicales qui opèrent la distinction sémantique dans la langue maternelle de l'annotateur. Pour le reste des exemples, il a été demandé aux annotateurs de fournir leur traduction préférée dans leur langue. Dans ce cas, l'estimation de la probabilité de

[107] Il est à souligner qu'Ide (*ibid*.) explore la lexicalisation dans des langues différentes des sens fournis par WordNet pour un ensemble de mots. Selon elle, le fait que les informations inter-langues puissent être utilisées pour déterminer des distinctions sémantiques indépendamment d'un ensemble pré-défini de sens (comme WordNet) n'est pas évident.

[108] Les conclusions déduites permettraient ainsi d'estimer la nécessité de considérer l'ensemble des langues ou seulement une partie (les langues proches ou distantes).

lexicalisation de la distinction sémantique dans cette langue est plus faible. La raison en est que les annotateurs ne sont pas confrontés à paires de mots et il ne leur est pas explicitement demandé d'utiliser des mots qui effectuent des distinctions sémantiques. Dans ces cas, le même mot peut être choisi pour deux sens présentant une différence subtile, même si une paire de mots pouvant capter cette différence existe. Par conséquent, cette mesure « capte » surtout la tendance à lexicaliser une distinction sémantique donnée par le choix lexical « préféré ». En revanche, les données utilisées dans le travail d'Ide sont des textes traduits par des traducteurs expérimentés, ayant accès à un contexte plus large lors du processus de traduction. Les équivalents étant extraits de textes traduits, les expériences d'Ide n'examinent pas si une distinction **peut** être lexicalisée dans une autre langue mais si elle **est** effectivement lexicalisée.

La position de Resnik et Yarowsky est également partagée par Dyvik (2003), qui se sert des relations de traduction entre les mots de deux langues afin de générer leurs représentations sémantiques. Selon Dyvik, les représentations sémantiques créées sur la base d'informations provenant de langues très proches[109] sont caractérisées par un degré de granularité très bas, qui s'explique par la haute fréquence de correspondances de type « un-à-un » entre les unités lexicales des deux langues. En revanche, les informations sémantiques provenant des relations de traduction entre langues typologiquement éloignées sont plus riches et, dans ces cas, la granularité des représentations sémantiques est alors beaucoup plus fine. La quantité d'informations sémantiques pouvant être acquises à partir des relations de traduction entre langues éloignées est, par conséquent, plus importante que la quantité d'informations issues de langues très proches.

Nous reviendrons sur cette notion de distance inter-langue dans le §1.1 du chapitre 8, traitant de la nécessité de désambiguïsation lexicale dans les systèmes de Traduction Automatique. Ce paramètre est considéré comme influençant de manière importante le besoin de désambiguïsation lexicale pour la traduction. Cependant, dans le cadre de la TA également, les positions défendues à l'égard de cette question s'avèrent contradictoires.

[109] L'exemple de langues très proches, utilisé par Dyvik, concerne la paire norvégien et danois.

Quels que soient le degré de lexicalisation et la granularité des sens lexicalisés dans d'autres langues, une question importante liée à la possibilité d'utilisation des informations de traduction pour l'analyse sémantique concerne la pertinence des distinctions sémantiques proposées par le biais d'autres langues. Il convient donc d'analyser désormais les hypothèses théoriques définissant la pertinence des distinctions sémantiques induites dans une langue à l'aide des informations de traduction.

2.4. Projection inter-langue d'informations sémantiques

2.4.1. Pertinence des distinctions sémantiques proposées

Les hypothèses théoriques sur lesquelles reposent les méthodes traductionnelles d'acquisition de sens concernant leur lexicalisation paraissent valables et les résultats fournis, malgré leurs divergences, sont encourageants quant à la possibilité d'utiliser les représentations sémantiques engendrées pour l'analyse sémantique.

Les distinctions sémantiques induites dans une langue par le biais d'une autre langue semblent adéquates au traitement réalisé dans le cadre de certaines applications multilingues. Néanmoins, sur un plan théorique, il serait souhaitable d'adopter une attitude plus réservée. Ce point est souligné par Fuchs (1996 : 86-87) : « … *on se gardera de considérer la traduction dans une langue-cible comme un* ***test*** *a posteriori de l'existence d'ambiguïtés dans la langue-source. Cette erreur revient à méconnaître la spécificité de chaque système linguistique, et à plaquer sur une forme de départ sémantiquement univoque (mais parfois relativement indéterminée) une distinction plus fine, ou simplement différente, opérée par une autre langue*. »

Les distinctions sémantiques présentes dans une langue cible et exprimées par les équivalents de traduction d'un mot source peuvent être valables uniquement au sein de cette langue. Dans ce cas, l'induction des distinctions sur les unités lexicales d'une LS, avec lesquelles elles entrent en relation de traduction, ne conduit pas à des distinctions sémantiques pertinentes au sein des

mots source[110]. Il est donc recommandé d'adopter une attitude réservée face à la projection de distinctions sémantiques d'une langue à une autre, et de proposer des critères supplémentaires, qui permettraient d'estimer la pertinence des distinctions proposées.

Une telle attitude de prudence face aux distinctions sémantiques proposées est également adoptée par Dyvik (2005). L'existence de plus d'une traduction possible pour un mot ne constitue pas, selon lui, un critère suffisant pour juger de l'ambiguïté du mot en question. Dyvik cite comme exemple deux ensembles de mots, en anglais et en allemand, qui constituent un champ sémantique précis. Cet exemple est illustré dans la figure 1. La correspondance traductionnelle du mot allemand *Hexe* aux deux mots anglais *hag* et *witch* n'implique pas que *Hexe* soit ambigu. La seule conclusion qui puisse être déduite de cette correspondance, sans risque d'erreur, est que la dénotation du mot allemand couvre celles des deux mots anglais. Si ambiguïté il y a dans ce cas, alors celle-ci doit être établie indépendamment, l'existence de plus d'une traduction n'étant pas une condition suffisante[111].

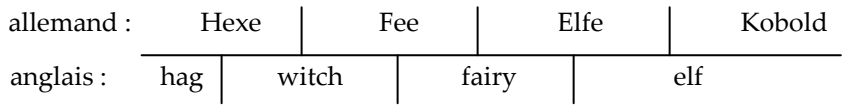


Figure 1. Correspondances entre mots anglais et allemands constituant un champ sémantique

[110] Des exemples cités par Fuchs (*ibid.*) pour illustrer ce point concernent les équivalents de traduction du mot français *mouton* en anglais : *sheep* et *mutton*, les mots utilisés par les esquimaux pour désigner la « neige » selon ses différents états (poudreuse, gelée, collante, etc.), ainsi que les mots utilisés en swahili pour désigner le « riz » (selon qu'il s'agit de la plante sur pied, de la céréale récoltée ou de l'aliment décortiqué que l'on fait cuire). L'existence de plus d'un équivalent de traduction, dans ces cas, ne démontre aucunement l'ambiguïté des mots français correspondants.

[111] Pour cette raison, dans la méthode des Miroirs Sémantiques proposée par Dyvik, les équivalents de traduction différents des mots polysémiques étudiés sont considérés comme des traits « contribuant » à proposer des sens au sein des mots polysémiques et non comme des indices clairs de distinctions sémantiques. Nous reviendrons plus en détail sur le fonctionnement de cette méthode dans la suite de ce chapitre.

2.4.2. Pertinence des informations extraites de corpus de traduction

Aux réserves déjà émises quant à la validité des distinctions sémantiques établies par projection inter-langue d'informations s'en ajoutent d'autres, lorsque ces informations sont repérées au sein de corpus de traduction (ou corpus parallèles). Les questions sur la nature des traductions, leur apport à une représentation de la langue ainsi que la pertinence des jugements pouvant en être déduits apparaissent régulièrement dans la littérature mais les réponses fournies divergent fortement. D'après Teubert (1996 : 247), les traductions donnent une fausse image de la langue qu'elles représentent et ce, même si elles sont de très bonne qualité, car elles constituent surtout un miroir de la LS. Les linguistes ne devraient donc pas se fier aux traductions pour la description d'une langue.

Cette méfiance à l'égard des traductions en tant que source de jugements linguistiques pertinents se manifeste par l'emploi de termes tels que **translationese** (Gellerstam, 1996 ; Shlesinger, 1992 ; Tirkkonen-Condit, 2002), **troisième code** et **langue hybride** (Frawley, 1984b). Le terme « translationese » sert à désigner la langue de la traduction, souvent considérée comme un type particulier de langue, marquée à la fois par l'influence de la LS et par certaines caractéristiques inhérentes au processus de traduction (Baker, 1993, 1995). Pour Frawley (*ibid*.) – le premier à avoir parlé d'une **langue de traduction** comme ayant une existence autonome –, la confrontation du texte source et de la LC lors du processus de traduction crée ce qu'il appelle un « troisième code ». Ce code, qui évolue pendant la traduction et dans lequel le texte cible est rédigé, constitue pour Frawley un compromis entre les normes et les structures de la LS et celles de la LC. Johansson (1998 : 6) remet aussi en question la possibilité de considérer les textes traduits comme représentatifs de l'usage ordinaire de la langue, en raison de l'influence de la LS et des traits généraux qui caractérisent les textes traduits. L'utilisation de corpus de traduction dans des études contrastives ne serait pas possible, d'après Johansson, si les effets du processus de traduction n'étaient pas contrôlés[112].

[112] Johansson (*ibid*.) fait des propositions en ce sens, comme l'inclusion dans le corpus de traductions dans les deux directions et l'examen d'observations tirées du corpus de traduction par référence à un corpus de « contrôle », constitué de textes originaux et traduits comparables dans la même langue, comparaison qui permettrait la validation des résultats de l'étude contrastive. Même

Des critiques contre cette manière de concevoir le produit de la traduction ont néanmoins été formulées. Pour Baker (1998), une analyse en profondeur de la notion de « troisième code » permettrait de rendre compte des particularités qui caractérisent la langue de traduction, en tant que phénomène distinct, d'une façon plus nuancée. Baker se prononce, en même temps, contre l'utilisation du terme « translationese » – qu'elle traduit en français par **jargon de traduction** – en raison de ses connotations péjoratives. Selon Baker, la traduction crée un troisième code parce qu'elle est une forme de communication unique, et non une forme de communication fautive, déviante ou contraire à la norme. Elle se lance, de son côté, vers la quête d'**universaux de traduction**, c'est-à-dire de traits qui caractérisent la langue traduite indépendamment des langues source et cible (Baker, 1993, 1995 ; Laviosa, 1998). Mauranen (2002) se prononce également contre l'utilisation du terme « translationese » par certains auteurs pour décrire un biais systématique observé dans la traduction. Elle se positionne tout autant contre la méfiance envers l'utilisation de traductions pour la description de la langue. Pour Mauranen, la langue de la traduction fait non seulement partie de la langue naturelle en usage mais devrait être traitée de manière adéquate et mériterait d'être étudiée.

Malgré les réserves exprimées quant à l'utilisation des traductions comme source de jugements linguistiques, la disponibilité croissante de corpus de traduction en fait une source importante d'informations pour les études de linguistique contrastive et de traduction. Nous avons déjà présenté, dans les paragraphes précédents, un ensemble de méthodes d'acquisition de sens, plus ou moins sophistiquées, qui exploitent les informations de traduction rencontrées dans des corpus parallèles. Nous avons aussi analysé certains facteurs qui influencent la réussite de ces méthodes et la qualité des résultats obtenus. La validation des résultats de l'analyse sémantique effectuée n'est pourtant pas facile, en raison du manque d'inventaires sémantiques de référence. Dans le paragraphe suivant, nous allons précisément aborder le problème de la

si ces résultats ne sont pas validés, ce processus permettrait néanmoins d'identifier des traits reflétant le processus de traduction ou les normes applicables aux textes traduits. Les déviations entre textes originaux et traductions dans la même langue pouvant être causées soit par les langues comparées soit par des caractéristiques des textes traduits en général, l'étude de traductions des mêmes textes source dans plusieurs langues permettrait de distinguer ce qui est spécifique aux langues de ce qui est propre à la traduction.

validation des résultats obtenus par les méthodes automatiques d'acquisition de sens, monolingues et traductionnelles.

3. Validation de sens automatiquement induits

3.1. Absence d'étalon d'or

L'absence d'un **étalon d'or** (*gold standard*) en sémantique lexicale, qui constituerait une solution aux problèmes du nombre et de la granularité des sens lexicaux, complique considérablement l'étape de la validation des distinctions sémantiques proposées par les méthodes automatiques d'acquisition de sens. Nous avons déjà analysé un ensemble de facteurs qui rendent difficile, voire impossible, la création d'une ressource sémantique unique utilisable par des humains aussi bien que par la machine, ou même adéquate à des applications automatiques différentes (cf. §1.3.2, chapitre 1). En outre, le manque d'uniformité entre ressources sémantiques différentes leur interdit d'être considérées comme une référence.

La difficulté à valider les résultats de l'acquisition de sens est encore accentuée par l'absence de **critères objectifs** d'estimation de la validité des sens proposés, problème lié tant aux divergences entre conceptions du sens qu'aux nombreux paramètres extra-linguistiques qui conditionnent l'adéquation des distinctions sémantiques au sein de cadres et d'applications différents.

3.2. Exploitation de ressources sémantiques externes

3.2.1. Inventaires sémantiques préétablis

Même si l'utilisation de ressources existantes en tant qu'étalon d'or est sévèrement critiquée, la comparaison des distinctions sémantiques obtenues automatiquement avec celles existant dans des ressources préétablies constitue un processus assez courant d'évaluation. Pantel et Lin (2002) emploient une évaluation de ce type, basée sur des correspondances automatiquement établies

entre les sens induits à partir d'un corpus et les sens des mots décrits dans le réseau sémantique WordNet (Miller *et al*., 1990)[113]. L'évaluation de la qualité des clusters sémantiques revient à estimer la qualité de ces correspondances. Purandare et Pedersen (2004b) valident les clusters obtenus à l'aide du même ensemble de sens[114].

Dans le travail d'Ide (1999a, 1999b), l'évaluation se fait également par référence à des ressources externes. L'acquisition de sens ne se produit cependant pas *ex nihilo*. Les sens obtenus correspondent à des clusters des sens fournis dans WordNet pour les mots étudiés. Les informations traductionnelles sont utilisées afin d'analyser la lexicalisation des sens de WordNet dans d'autres langues, processus qui peut également servir à estimer la validité des sens en question. Plus précisément, les informations de traduction servent à estimer la similarité des sens de WordNet, similarité qui sert, à son tour, à leur clustering[115]. Les clusters de sens obtenus automatiquement par ce processus sont ensuite comparés aux entrées correspondantes dans des dictionnaires[116].

La difficulté à comparer les clusters de sens et les entrées dictionnairiques est néanmoins soulignée : hormis certaines similarités observées à un niveau grossier de distinctions, rendues évidentes par les clusters de sens générés automatiquement et les sens principaux décrits dans les entrées dictionnairiques, un taux élevé de divergence est constaté au niveau des distinctions de granularité fine. Plus précisément, les sens de granularité fine sont

[113] A l'intérieur de ce réseau, les sens sont représentés à l'aide de **synsets**, ensembles de mots synonymes désignant un concept lexical. Les synsets sont liés dans le réseau par des relations sémantiques, comme l'hyponymie et l'hypéronymie, l'antonymie ou la méronymie.

[114] Les données de SENSEVAL-2, utilisées pour l'évaluation, sont sémantiquement étiquetées à l'aide des sens de WordNet (Kilgarriff, 2001).

[115] Un indice de cohérence est calculé qui mesure la tendance de lexicalisation différente des sens différents des mots source décrits dans WordNet, dans un certain nombre de langues. Les occurrences des mots polysémiques étudiés au sein du corpus parallèle sont regroupées à la main en fonction des distinctions sémantiques de WordNet ; leurs traductions utilisées dans le corpus sont repérées. L'indice de cohérence constitue une indication du degré de validité des distinctions sémantiques et démontre la similarité des sens correspondants : plus la valeur de l'indice entre deux sens est grande, et plus les occurrences des sens en question sont traduites fréquemment par la même unité lexicale dans les autres langues. Une tentative de validation des sens décrits dans un autre inventaire sémantique (l'inventaire HECTOR, utilisé dans la campagne d'évaluation SENSEVAL), par utilisation d'informations de lexicalisation inter-langue, est aussi entreprise par Resnik et Yarowsky (2000).

[116] Quatre dictionnaires sont utilisés : Colling English (CED), Longman's (LDOCE), Oxford Advances Learner's Dictionary (OALD) et COBUILD.

soit éparpillés à différents endroits dans les entrées dictionnairiques soit partagés entre sens différents. Les relations hiérarchiques évidentes dans les clusters générés par l'algorithme hiérarchique de clustering ne sont pas non plus reflétées dans les entrées dictionnairiques, où les sens sont présentés de manière linéaire.

La difficulté à comparer les résultats fournis par une méthode automatique d'acquisition de sens aux informations contenues dans des ressources sémantiques préétablies est également soulignée par Thunes (2003), lors d'une évaluation de la qualité des résultats de la méthode des Miroirs Sémantiques (Dyvik, 2003, 2005). Dans ce travail d'évaluation, les résultats d'analyse sémantique fournis pour un mot par la méthode des Miroirs sont comparés aux informations trouvées, pour ce même mot, dans les entrées de WordNet et du Merriam-Webster Online Thesaurus. Les résultats de l'évaluation quantitative initialement effectuée sont complétés et réajustés par une évaluation qualitative, menée sur la base de descriptions des sens et des sous-sens dans les deux ressources de référence.

L'évaluation quantitative s'effectue par le calcul de la précision et du rappel, en considérant des ensembles de lemmes répertoriés en tant que mots sémantiquement similaires du mot polysémique, dans chacune des trois entrées : l'entrée correspondante au mot dans la ressource générée par les Miroirs et les entrées correspondantes de WordNet et de Merriam Webster.

La **précision** est calculée comme étant la proportion de mots partagés entre l'entrée générée par les Miroirs et les entrées de l'étalon d'or. Des ajustements qualitatifs ont lieu par la suite, en se référant aux descriptions sémantiques fournies dans les entrées de l'étalon d'or et des nuances de sens couvertes par les différents ensembles de mots sémantiquement proches. Ainsi, il est estimé que des mots apparaissant uniquement dans l'entrée des Miroirs, pour le mot étudié, auraient pu être inclus dans les entrées de l'étalon d'or[117]. Le **rappel** indique, quant à lui, la proportion des mots contenus dans les ressources qui ne sont pas repérés par la méthode des Miroirs. L'ajustement qualitatif entrepris dans le cas du rappel consiste à vérifier la présence des mots qui sont absents de l'entrée des Miroirs, au sein du corpus d'apprentissage. L'absence de ces mots provoque un

[117] Pendant cette étape, il faut examiner si les mots qui apparaissent seulement dans l'entrée des Miroirs pourraient être pris en compte par la description sémantique donnée dans l'entrée de l'étalon d'or. Si c'est le cas, les mots en question auraient pu être inclus dans l'étalon d'or.

réajustement du rappel[118]. Après ce réajustement, le rappel reflète la proportion des mots contenus dans les entrées des ressources de référence qui « pourraient » être trouvés et qui ont effectivement été trouvés par la méthode. Enfin, l'évaluation qualitative examine également si les mots absents de l'entrée des Miroirs introduiraient des nuances de sens qui ne sont pas décrites par l'ensemble des mots inclus dans l'entrée en question. La dernière étape de l'évaluation comporte une comparaison qualitative entre les distinctions sémantiques (sens et sous-sens) présentes dans l'entrée des Miroirs et celles décrites au sein des entrées de l'étalon d'or[119].

L'évaluation de la description sémantique fournie par la méthode des Miroirs pour un adjectif anglais (le mot *pleasant*) démontre le faible nombre de mots communs aux trois entrées correspondantes à ce mot. Les entrées des Miroirs et de Webster partagent à peu près le même nombre de mots avec l'entrée de WordNet. Ces entrées fournissent par ailleurs à peu près le même nombre total de mots pour le mot étudié, nombre qui est supérieur au double du nombre de mots trouvés dans l'entrée de WordNet. Autre élément très intéressant, de grandes divergences existent au niveau des entrées trouvées dans les deux ressources préexistantes, ce qui remet en question leur statut en tant qu'étalon d'or.

Le recours à des ressources prédéfinies pour l'évaluation s'explique en partie par la difficulté pratique et la subjectivité inhérentes à une **évaluation manuelle** des sens induits par le corpus (Véronis, 2004). D'après Agirre *et al*. (2006) et Agirre et Soroa (2007a), une évaluation de ce type pourrait impliquer l'estimation de la validité d'attribution des sens obtenus pour un mot à ses instances dans un corpus. Hormis la difficulté pratique que représente la vérification manuelle de chaque instance, décider de la conformité de la

[118] Ce réajustement a lieu dans le cas où un mot apparaît dans le corpus mais n'est jamais traduit par un équivalent de traduction identifiable dans le texte parallèle ; ce qui explique que la méthode des Miroirs ne puisse pas le repérer. En outre, dans le cas où un mot est toujours traduit par le même équivalent, la méthode ne peut pas décider de sa proximité sémantique avec d'autres mots de la langue, ce qui explique qu'il soit absent de l'entrée générée. Pourtant, le rappel n'est pas ajusté dans ces cas, car l'absence de ces mots est due à la manière dont la méthode est conçue.

[119] Dans la mesure où il y a rarement une solution unique quant à la division du sens d'un mot en sens et en sous-sens (Dyvik, 2003), la méthode des Miroirs offre la possibilité de modifier la granularité des distinctions sémantiques, en modifiant la valeur d'un paramètre, appelé 'Seuil de Recouvrement' ('Overlap Threshold'). Pour l'évaluation menée par Thunes, la valeur de ce paramètre était la valeur par défaut.

correspondance d'une instance au sens décrit par le cluster[120] qui lui est attribué s'avère également difficile, surtout lorsque le cluster comprend un petit nombre de mots. Dans ce cas, le sujet, au lieu d'estimer la justesse du cluster fourni par l'algorithme, étiquetterait l'instance avec ses propres sens, sens ensuite comparés au cluster fourni par le système.

3.2.2. Corpus sémantiquement étiquetés

Le problème de la **subjectivité** qu'implique une évaluation manuelle peut être résolu en comparant le résultat du clustering aux informations provenant d'un corpus étiqueté avec des sens de référence (étalon d'or). Ce type d'évaluation est adopté dans SemEval (Agirre et Soroa, 2007a) pour la tâche qui consiste à évaluer les résultats de systèmes de repérage de sens. Les sens de référence proviennent d'une ressource préétablie, le réseau 'OntoNotes'[121] (Hovy *et al*., 2006), dont les distinctions sémantiques sont de granularité plus grossière que WordNet. Lors d'une étape d'**évaluation non supervisée**, les sens induits à partir d'un corpus non étiqueté sont considérés comme des clusters d'exemples et les sens de l'étalon d'or comme des « classes ». Les clusters sont alors comparés avec les ensembles d'exemples étiquetés par les sens de l'étalon d'or (classes). Un clustering parfait correspondrait à l'état où chaque cluster inclut exactement les mêmes exemples qu'une classe, et réciproquement.

SemEval[122] comprend également une étape d'**évaluation supervisée**, concernant la mise en correspondance des sens induits avec les sens de l'étalon d'or, et l'utilisation de ces correspondances pour étiqueter le corpus de test avec les étiquettes de l'étalon d'or. Dans ce cas, le corpus est divisé en deux, une partie d'entraînement et une partie de test. La correspondance entre clusters et sens est calculée en utilisant les informations sémantiques d'annotation dans la partie

[120] Le sens peut être décrit autrement, par exemple par une sous-partie d'un graphe, comme chez Véronis (2004).

[121] Cependant Agirre et Soroa (*ibid*.) soulignent que l'utilisation de WordNet, au lieu d'OntoNotes, aurait donné de meilleurs résultats.

[122] Chaque mesure utilisée pour évaluer le clustering est orientée en fonction d'une certaine stratégie de clustering. La mesure utilisée dans SemEval pénalise davantage les systèmes donnant un grand nombre de clusters tandis qu'elle favorise ceux induisant moins de sens. L'évaluation supervisée semble être plus « neutre » en ce qui concerne le nombre de clusters.

d'entraînement. Les résultats sont ensuite évalués à l'aide des mesures de précision et de rappel, employées pour l'évaluation des systèmes de désambiguïsation lexicale supervisée.

3.3. Exploitation de sens induits en vue de tâches précises

Il existe une manière alternative de valider le contenu d'une ressource sémantique, générée automatiquement ou non, et qui ne nécessite ni la référence à des ressources préétablies ni l'utilisation de corpus sémantiquement étiquetés. Cette méthode consiste à estimer la possibilité d'utiliser cette ressource sémantique pour une tâche précise. Tufiş *et al*. (2004b) évaluent la qualité des contenus de BalkaNet en les exploitant dans une tâche de désambiguïsation lexicale. Le système de désambiguïsation développé dans le cadre de ce travail est basé sur des corpus parallèles et exploite l'intuition selon laquelle les mots qui sont des traductions réciproques dans des textes parallèles devraient avoir les mêmes sens inter-langues (ou similaires) au sein de la ressource[123]. Ainsi, le système de désambiguïsation fait office d'outil de validation, en permettant aussi bien le repérage d'alignements erronés entre les wordnets de langues différentes, que le repérage de synsets incomplets ou manquants.

Le travail de Lyse (2006) vise à valider les résultats obtenus par la méthode des Miroirs Sémantiques[124], en estimant leur utilité en tant que source de connaissances lexicales dans le cadre d'une tâche de désambiguïsation lexicale supervisée[125]. Le corpus d'apprentissage utilisé est sémantiquement étiqueté à l'aide des résultats des Miroirs, ce qui permet l' « enrichissement » du contexte par des informations paradigmatiques. L'apprentissage porte ensuite sur des

[123] Le réseau BalkaNet comprend les wordnets construits pour six langues différentes, alignés au Princeton WordNet considéré comme un index inter-langue. La méthode de désambiguïsation qui exploite ce réseau sera décrite plus en détail dans le paragraphe 2.5.2.3.

[124] Il s'agit des résultats obtenus par application de la méthode des Miroirs Sémantiques sur un corpus parallèle anglais-norvégien (English-Norwegian Parallel Corpus, ENPC) de 2.6 millions de mots. Les résultats des Miroirs constituent un inventaire sémantique, qui décrit des relations paradigmatiques entre les mots (relations de quasi-synonymie, d'hyperonymie et d'hyponymie).

[125] La désambiguïsation lexicale par apprentissage automatique supervisé nécessite un corpus d'apprentissage comprenant les exemples de contextes dans lesquels les sens apparaissent. Chaque occurrence des mots polysémiques dans ce corpus doit être étiquetée par son sens correct avant l'apprentissage ; ainsi le système « apprend » ce qui caractérise le contexte d'un sens donné et le classificateur qui en résulte peut ensuite être utilisé pour classifier de nouvelles occurrences des mots polysémiques.

classes de sens similaires présentes dans le contexte et non sur les mots du contexte. Ces classes sont constituées à l'aide des traits sémantiques attribués aux sens par la méthode des Miroirs Sémantiques[126]. La précision de l'étiqueteur sémantique est haute et la comparaison des résultats de cette expérience avec ceux d'une tâche de désambiguïsation où l'apprentissage est basé sur les mots du contexte a démontré que l'apprentissage sur les traits sémantiques des mots (même si les mots ne sont que les noms du contexte) donne de meilleurs résultats. Dans le travail de Schütze (1998), les clusters de sens obtenus par la méthode non supervisée sont utilisés pour la recherche d'information. Les inconvénients de ce type d'évaluation résident, selon Agirre et Soroa (2007a) et Agirre *et al*. (2006), dans le besoin de développer des systèmes appropriés, ce qui n'est pas toujours évident, et dans la difficulté à expliquer une bonne ou une mauvaise performance.

3.4. Validation des sens induits au sein de ce travail

En ce qui concerne notre démarche, nous avons opté pour une autre manière de valider les résultats de la méthode d'acquisition de sens. Cette validation se fait par comparaison des résultats obtenus par la méthode proposée, sur un échantillon de mots ambigus, à ceux fournis, sur le même échantillon, par une méthode qui exploite des informations de nature différente pour l'acquisition de sens, la méthode des **Miroirs Sémantiques**. La similarité des résultats de ces deux méthodes, basées sur des principes différents, ne peut être considérée comme un effet du hasard et, par conséquent, servira d'appui aux distinctions sémantiques proposées.

Les descriptions sémantiques obtenues seront également comparées à celles fournies par le réseau sémantique multilingue **BalkaNet**. Cette comparaison aura comme objectif principal de démontrer les différences au niveau des descriptions sémantiques entre la ressource automatiquement générée et cette ressource

[126] Les traits sémantiques constituent un moyen formel de représentation de la similarité entre les sens lexicaux : les sens sont d'autant plus proches que le nombre de traits sémantiques qu'ils partagent est élevé. Ainsi des mots *lunch* et *diner*, qui sont des hyponymes de *meal*, relation exprimée par le partage d'un même trait sémantique par les trois sens (les hyponymes étant caractérisés par un trait supplémentaire, propre à chacun, qui les distingue). Pour plus de détails sur la méthode, voir Lyse (2006).

prédéfinie, et de mettre en évidence les qualités des descriptions engendrées par notre méthode. Cette comparaison se heurte aux difficultés soulignées lors des tentatives de comparaison des résultats des méthodes automatiques aux contenus de ressources préétablies. Elle permet, néanmoins, d'avoir une image tant des divergences qualitatives existantes au niveau des descriptions sémantiques, que de celles concernant la structure des inventaires en question.

CONCLUSION

Dans ce chapitre, nous avons présenté un ensemble de méthodes automatiques d'acquisition de sens, opératoires dans un cadre monolingue et bi-(multi-)lingue. Nous avons analysé les hypothèses théoriques sous-jacentes à ces méthodes, ainsi que les principes de leur fonctionnement et les facteurs conditionnant leur réussite. Nous avons également décrit la nature des résultats fournis, c'est-à-dire la nature des informations contenues dans les inventaires de sens construits. Nous avons enfin souligné les difficultés rencontrées lors des tentatives d'évaluation des résultats obtenus.

L'un des usages principaux des inventaires sémantiques construits à l'aide de méthodes automatiques d'acquisition de sens est de fournir l'ensemble des sens nécessaire à la sélection du sens de nouvelles instances de mots, lors des tâches de désambiguïsation lexicale. Les résultats de la méthode d'acquisition de sens proposée dans ce travail seront aussi exploités pour la désambiguïsation et pour la sélection lexicale dans le cadre de la traduction. Dans le chapitre suivant, nous allons analyser quelques propositions liées à la résolution de la polysémie dans un cadre automatique et décrire certains aspects intéressants concernant le fonctionnement des méthodes de désambiguïsation.

CHAPITRE 3

DESAMBIGUÏSATION LEXICALE

INTRODUCTION

Le processus de désambiguïsation lexicale (Word Sense Disambiguation – WSD) consiste à sélectionner les sens corrects d'instances contextualisées des mots ambigus, parmi l'ensemble de leurs sens possibles (ou **sens candidats**). Cette sélection présuppose donc l'existence d'un inventaire de sens lexicaux. Les sens choisis à l'issue de l'étape de désambiguïsation peuvent être directement exploités pour une tâche précise au sein d'une application (par ex. la traduction des mots dans une autre langue) ; ils peuvent aussi servir de métadonnées pour l'étiquetage sémantique de textes, ce qui permet la création de ressources enrichies par des informations sémantiques.

La désambiguïsation s'effectue en mettant en correspondance, à l'aide d'une méthode d'association, les informations relatives aux nouvelles instances des

mots ambigus avec celles provenant d'une source externe (désambiguïsation basée sur les connaissances) ou avec des informations trouvées dans des corpus textuels (désambiguïsation dirigée par les données). L'avantage des méthodes du deuxième type réside en ce que les informations requises ne nécessitent pas une modélisation étendue, contrairement aux informations (lexico-sémantiques, encyclopédiques ou autres) requises par les méthodes basées sur les connaissances[127]. Les méthodes dirigées par les données se divisent, quant à elles, en méthodes supervisées et méthodes non-supervisées ; les premières s'appuient sur un ensemble d'apprentissage réunissant des exemples d'instances désambiguïsées des mots, tandis que les deuxièmes exploitent les résultats de méthodes automatiques d'acquisition de sens (cf. chapitre 2).

Nous allons, dans ce chapitre, analyser la nature des informations pouvant contribuer à la désambiguïsation et les principes de fonctionnement des différents types de méthodes de désambiguïsation lexicale.

1. Informations exploitées pour la levée de l'ambiguïté

1.1. Le rôle du contexte dans la désambiguïsation

La seule manière d'identifier le sens d'un mot ambigu est de se référer à son contexte (Ide et Véronis, 1998), ce qui explique la place prépondérante du contexte dans tous les travaux de désambiguïsation. Plus précisément, la sélection du sens correct véhiculé par une nouvelle instance d'un mot ambigu consiste à supprimer les acceptions qui génèrent un « conflit » sémantique avec le contexte de l'instance. A la fin de ce processus, si toutes les acceptions sont supprimées sauf une, celle-ci est sélectionnée et attribuée à l'instance en question[128]. Ce mécanisme de sélection est caractéristique des méthodes

[127] Une telle modélisation n'est pas évidente pour tous les domaines de connaissance ni pour un grand nombre de langues.

[128] Dans le cadre de la communication humaine, le contexte peut avoir, outre la sélection, d'autres effets sur le sens d'un mot, comme la **coercition** et la **modulation** (Cruse, 2004 : 118-120). Dans le cas de la coercition, si aucune des acceptions établies d'un mot n'est compatible avec le contexte, les récepteurs du message cherchent des extensions possibles de sens (comme la métaphore et la métonymie), afin de trouver une acception compatible avec le contexte (en admettant l'hypothèse

automatiques de désambiguïsation. Il ne faut bien évidemment pas s'attendre à retrouver un tel fonctionnement dans le cadre de la communication humaine ; s'il peut se produire, consciemment ou inconsciemment, il est fort probable que les alternatives d'interprétation ne franchissent pas le seuil de conscience du locuteur et que la sélection du sens correct soit effectuée sans considérer les autres possibilités (Bréal, 1899 : 156).

Le contexte est donc conçu comme ayant, le plus souvent, un **rôle réducteur** : il ressemble à un « filtre » qui permet de lever nombre d'ambiguïtés virtuelles. Ce rôle du contexte dépend des mots qui y apparaissent. L'inclusion dans le contexte de mots sémantiquement apparentés à l'un des sens du mot ambigu (également appelés **mots amorces**) facilite la sélection de ce sens parmi l'ensemble des sens possibles du mot[129]. Le contexte peut, en outre, être caractérisé comme **inducteur**, dans le cas où il existe une **affinité préférentielle** entre lui et l'une des significations de l'expression ambiguë, affinité que l'on peut traduire en termes de probabilité relative d'apparition de la significations en présence du contexte considéré (Fuchs, 1996 : 59)[130].

Une remarque générale s'impose, à propos du rôle du contexte dans la levée de l'ambiguïté : la désambiguïsation de mots caractérisés par des types ou des

que les locuteurs essaient de transmettre un message intelligible). Si une telle acception est trouvée, elle est alors considérée comme l'acception visée et le contexte est considéré comme ayant induit une nouvelle acception. En revanche, la modulation peut être perçue comme de la variation contextuelle qui ne traverse pas les limites d'un seul sens et englobe des phénomènes comme l'**enrichissement** et l' **appauvrissement**, qui dépendent du type de l'effet exercé par le contexte sur le mot. L'enrichissement ajoute du contenu sémantique, il enrichit le sens ou le rend plus spécifique (spécialisation hyponymique ou méronymique). L'appauvrissement survient lorsque le contexte rend évidente l'utilisation du mot dans un sens vague (Cruse, *ibid*.). Ainsi, le contexte ne lève pas toujours les ambiguïtés mais peut en révéler, voire même en créer de nouvelles (Fuchs, 1996 : 53-55).

[129] Le rôle important des « mots amorces » sur la sélection du sens correct d'une instance d'un mot polysémique a été démontré par des expériences menées en psycholinguistique, qui mesurent la facilité avec laquelle le sujet sélectionne le sens correct du mot en présence de mots amorces situés à proximité (Kintsch et Mross, 1985). Les résultats de ces expériences ont montré que les collocations sont traitées de façon différente des autres cooccurrences : les mots amorces qui se trouvent en collocation fréquente avec les mots polysémiques servent à les activer dans les tâches de décision lexicale, tandis que ceux qui sont liés au contexte thématique ne facilitent pas les décisions lexicales des sujets.

[130] Comme nous avons déjà souligné, l'étude de la notion de « contexte » dans notre travail se limite au contexte linguistique. Nous pouvons néanmoins préciser que les facteurs extra-linguistiques ont, à l'instar du contexte linguistique, un rôle à la fois réducteur et démultiplicateur en matière d'interprétation. Ces facteurs exercent en effet une action déterminante dans l'interprétation, qui explique que peu d'ambiguïtés linguistiques effectives donnent lieu à de réelles équivoques en situation (Fuchs, 1996 : 61).

degrés différents d'ambiguïté ne requiert pas le même type d'informations. La nature et la quantité des informations contextuelles requises sont dépendantes des relations entretenues entre les différents sens des mots, de leur distinctivité et de leur exclusion. La résolution de l'**ambiguïté contrastive** (Weinreich, 1964 ; Pustejovsky, 1996 : 2), par exemple, est considérée dans la plupart des travaux de désambiguïsation comme un cas relativement simple. Les sens contrastifs sont en effet souvent de **nature exclusive**, c'est-à-dire qu'un sens n'est disponible que si aucun autre ne l'est dans un contexte donné. Pour pouvoir opérer une sélection entre sens différents d'une unité lexicale caractérisée par ce type d'ambiguïté, la prise en compte d'informations sur le domaine ou le sujet traité peut donc s'avérer suffisante.

1.2. Exploitation des informations du domaine

1.2.1. Levée de l'ambiguïté par restriction à des domaines précis

Le rôle du **domaine** comme paramètre important de la restriction de l'ambiguïté lexicale a déjà été reconnu dans les premiers travaux de désambiguïsation menés dans le cadre de la TA. La solution proposée pour la réduction des sens possibles des mots ambigus – et la sélection de traductions pour de nouvelles instances de ces mots dans les textes – consistait à construire des **micro-glossaires**, c'est-à-dire des glossaires destinés à n'être utilisés qu'au sein de domaines spécialisés (Oswald, 1952 ; Reifler, 1954). Les sens des mots ambigus décrits dans ce type de glossaire étaient réduits aux sens pertinents dans le domaine concerné, ce qui éliminait une partie de son ambiguïté, proportionnelle à la spécialisation du domaine.

L'impact des informations du domaine et du sujet traité dans la désambiguïsation est en effet si important que, dans certains cas, ces informations suffisent à elles seules à sélectionner le sens correct des mots. Même si elles ne sont pas fournies au sein de ressources spécialisées, comme les micro-glossaires, elles peuvent être néanmoins repérées à l'échelle du document ou dans des portions de texte plus petites. Pour chaque sujet, il existe un sous-

vocabulaire de termes appropriés le désignant. Les méthodes de désambiguïsation qui utilisent ce type d'informations contextuelles exploitent la **redondance** dans les textes, c'est-à-dire l'usage répétitif de mots sémantiquement liés à un sujet précis. Dans ces cas, le contexte est traité le plus souvent comme un **sac de mots**, autrement dit un ensemble de mots non ordonné ; ce qui importe étant la cooccurrence d'un sens précis du mot ambigu avec des mots liés au sujet traité au sein d'une fenêtre textuelle. La tâche de désambiguïsation consiste alors à identifier le sujet traité par le nouveau texte et à sélectionner le sens du mot ambigu le plus adapté.

Dans un nombre important de travaux, le domaine traité constitue donc la principale source d'informations pour la désambiguïsation. Gale *et al*. (1992a) proposent de recourir à des méthodes proches des méthodes de **Recherche d'Information**[131]. Les informations exploitées par ce type de méthodes, lors des phases d'entraînement (phase servant à élaborer des discriminateurs de sens) et d'évaluation (étape de désambiguïsation de nouvelles occurrences des mots ambigus), sont repérées dans des contextes très larges (100 mots autour du mot ambigu). La désambiguïsation se base sur le principe **un sens par discours**, principe régissant également la méthode proposée par Yarowsky (1995) [132], selon lequel, le sens d'un mot est le même tout au long d'un document. Autrement dit, les différentes instances du mot dans le texte véhiculent tous le même sens. Les deux méthodes précitées (Gale *et al*, *ibid*., Yarowsky, *ibid*.) visent néanmoins à distinguer les deux sens principaux des mots étudiés, liés à des sujets différents[133].

En revanche, le travail de Magnini et Cavaglià (2000) s'attache aux sens des mots décrits dans WordNet. Les synsets de WordNet sont étiquetés à l'aide d'informations de domaine (Magnini et Cavaglià, *ibid*.) et Magnini *et al*. (2002)

[131] Les méthodes de recherche d'information visent la discrimination entre documents en fonction du sujet traité et l'identification de nouveaux documents relatifs à un sujet donné.

[132] Dans le travail de Yarowsky, cette hypothèse est combinée avec le principe **un sens par collocation**. Ce principe prend en compte le contexte local du mot à désambiguïser, supposé fournir des indices forts et consistants sur le sens du mot, conditionnés par la distance relative, l'ordre et la relation syntaxique (Yarowsky, 1995). Le mot « collocation » est employé ici dans son sens traditionnel, à savoir, les mots apparaissant au même endroit ou une juxtaposition de mots. Aucune interprétation idiomatique ou non-compositionnelle n'y est impliquée.

[133] Par exemple, le sens judiciaire et le sens grammatical du mot anglais *sentence* (Gale *et al*., 1992), et les sens animal et musical du mot anglais *bass* (Yarowsky, 1995).

utilisent ensuite ces informations pour la désambiguïsation[134]. Ces informations permettent l'établissement de relations entre les sens lexicaux, qui peuvent être utilisées de manière profitable lors du processus de désambiguïsation. Des vecteurs construits à partir de WordNet déterminent les domaines pertinents pour les sens des mots ambigus et sont, par la suite, comparés avec les vecteurs construits sur la base des contextes des nouvelles instances de mots. Le sens correspondant au vecteur le plus proche au vecteur du contexte est alors sélectionné comme étant le sens approprié du mot[135].

1.2.2. Limites de l'apport du domaine pour la désambiguïsation

La caractérisation d'un texte par rapport au domaine et au sujet traité peut donc, dans certains cas, aider à la désambiguïsation lexicale de manière importante. Ce processus n'est pourtant pas aussi simple qu'il le paraît, étant donné le degré variable de spécialisation des textes[136] et l'existence possible de sujets différents à l'intérieur d'un texte, et ce, même au niveau de petites sections textuelles. Une autre source de complication, entravant le bon fonctionnement de ces méthodes, concerne le nombre de sens d'un mot qui sont liés à un domaine ; lorsque ce nombre est supérieur à un, les informations du domaine ne suffisent évidemment pas pour pouvoir choisir entre les sens.

Cet aspect est souligné par Sparck Jones (1986 : 15-18), qui considère les étiquettes de domaine comme trop « grossières » et non pertinentes pour la description du sens d'un mot dans un contexte particulier : un mot peut en effet être utilisé dans des sens différents dans un texte qui traite pourtant d'un sujet bien précis[137]. Yarowsky (1992), en exploitant les informations liées aux catégories

[134] Les informations de domaine permettent l'établissement de relations entre les sens décrits dans WordNet, et rendent possible leur regroupement et, par conséquent, la réduction de leur granularité (souvent critiquée comme étant trop fine).

[135] Les résultats sont assez faibles du point de vue du rappel ; la raison en est, d'après les auteurs, que les mots des contextes, utilisés pour la désambiguïsation, ne véhiculent pas suffisamment d'informations relatives aux domaines.

[136] Le terme « texte spécialisé » peut être appliqué à des textes caractérisés par des degrés de spécialisation différents. L'ensemble de sens possibles des mots ambigus au sein d'un texte est d'autant plus petit que la spécialisation du texte est grande.

[137] Nous verrons plus bas l'exemple de l'homonyme anglais *plant*, dont les deux sens contrastifs, « plante » et « usine », se manifestent au sein de textes relatifs à la protection de l'environnement.

sémantiques du thésaurus Roget's, souligne également les limites d'une telle méthode de désambiguïsation. Ces limites caractérisent, d'une part, les mots qui présentent des distinctions sémantiques indépendantes d'un sujet précis et, d'autre part, les cas où des distinctions sémantiques fines peuvent être repérées au sein d'une catégorie du thésaurus[138].

La prise en compte des informations de domaine ne permet que le traitement de certains cas particuliers d'ambiguïté. Gale *et al*. (1992a, 1993) soulignent d'ailleurs que les exemples utilisés pour illustrer le fonctionnement des méthodes de ce type sont bien choisis et concernent des mots particulièrement adaptés à une telle procédure de désambiguïsation, dans la mesure où leur contexte contient souvent des indices très forts. La désambiguïsation à l'aide d'informations relatives au **domaine** et au **sujet traité** réussit effectivement bien surtout dans le cas de **sens lexicaux bien distincts**. Les informations de ce type ne sont pas suffisantes pour distinguer et sélectionner des sens moins clairement distincts et plus apparentés. Par exemple, dans le cas de la polysémie logique (Pustejovsky, 1995), où les sens d'un mot sont complémentaires et non exclusifs, et où ils ont un effet d' « ombrage » beaucoup plus faible les uns sur les autres, la désambiguïsation par prise en compte d'informations de domaine n'est pas possible. Les sens différents d'un mot à polysémie logique peuvent paraître tous équivalents pour l'interprétation du mot dans un domaine précis, bien qu'un seul sens soit visé dans un contexte particulier.

La sémantique des mots traités influence donc fortement l'**applicabilité** et l'**efficacité** des méthodes de désambiguïsation basées sur les informations de domaine. Des divergences quant à l'efficacité de ces méthodes peuvent aussi être observées au niveau d'un seul mot ; ainsi lorsque le mot est caractérisé conjointement par homonymie et polysémie[139].

Les informations de domaine s'avèrent également insuffisantes pour désambiguïser des mots dont les différents sens n'ont pas de lien clair à un sujet précis (Sparck Jones, 1986 : 15-18 ; Yarowsky, 1992 ; Leacock *et al*., 1998).

[138] Telle est, par exemple, la distinction entre les sens médical et narcotique du mot *drug*, qui sont regroupés au sein de la catégorie 'REMEDY' du thésaurus Roget's.

[139] Les sens contrastifs sont facilement distinguables à l'aide d'informations relatives au domaine et au sujet traité, ce qui n'est pas le cas pour les autres sens du mot, de granularité plus fine.

L'existence de sens qui ne sont pas limités à des sujets particuliers mais qui, au contraire, apparaissent librement dans des domaines différents de discours, fixe d'autres limites à l'applicabilité du principe « un sens par discours » de Gale *et al*. (1992b).

1.3. Le contexte local ou « micro-contexte »

1.3.1. Taille du contexte

L'inadéquation constatée des informations de domaine à la désambiguïsation dans un grand nombre de cas a généré la recherche d'autres sources d'informations plus appropriées, dont la plus importante est le **contexte lexical** ou **local** des mots[140]. Le contexte local (ou **micro-contexte**) d'un mot concerne les mots qui apparaissent à proximité de ce mot dans le texte. Selon Weaver (1949), si l'on examine séparément chacun des mots dans un livre, comme à travers un masque opaque avec une fente de la taille d'un mot, il est alors impossible de déterminer leur sens. Cependant, si on élargit la fente du masque, jusqu'à ce que l'on puisse voir non seulement le mot cible en question mais aussi *N* mots de chaque côté, alors si *N* est assez grand, on peut décider de manière non ambiguë du sens du mot cible.

Le contexte local est donc souvent délimité à l'aide d'une **fenêtre textuelle** qui se situe à gauche ou à droite ou des deux côtés d'une instance du mot ambigu et dont la taille peut varier[141]. La définition de la taille de la fenêtre textuelle est liée à celle de la **distance optimale** entre le mot ambigu et les indices contextuels pouvant servir à sa désambiguïsation. La détermination de cette distance optimale a fait l'objet d'un grand nombre de travaux dont les résultats sont assez variés.

[140] Le contexte lexical est aussi appelé co-texte (Fuchs, 1994 : 133), ce qui permet de le distinguer du contexte plus général qui entoure l'acte de communication, et qui peut s'appliquer aussi bien aux informations venant du texte qu'aux paramètres extra-linguistiques de communication (les conditions spatiales et temporelles, les participants à l'acte de communication, les connaissances du domaine etc.).

[141] Les fenêtres peuvent être soit délimitées à l'aide de séparateurs de phrases ou de paragraphes, soit définies à l'aide de « *n*-grammes », qui permettent l'observation d'un certain nombre (*n*-1) de mots entourant le mot polysémique dans le texte.

Pour Kaplan (1955), par exemple, le mot précédant le mot polysémique dans le texte est un très mauvais indice de désambiguïsation et nettement moins approprié que le mot suivant. Une fenêtre comprenant un mot de chaque côté du mot polysémique est plus efficace que celle qui en contient deux, et l'intérêt de retenir deux mots de chaque côté du mot polysémique est comparable à celui de la phrase entière. La pertinence des contextes très limités (± 1 ou ± 2 mots autour du mot polysémique) est également défendue par Choueka et Lusignan (1985) mais essentiellement pour la désambiguïsation des homographes[142]. Leacock *et al*. (1998) utilisent une fenêtre de ± 3 mots autour du mot ambigu, tandis que le classificateur utilisé par Bruce et Wiebe (1994a, 1994b) prend en compte ± 2 mots autour du mot ambigu. Pour Yarowsky (1993, 1994), la taille optimale de la fenêtre textuelle dépend du **type de l'ambiguïté** qui caractérise les mots : une grande fenêtre (de 20 à 50 mots) autour du mot ambigu est considérée comme étant optimale pour les ambiguïtés sémantiques ou relatives au sujet traité, tandis que pour les ambiguïtés syntaxiques locales, une plus petite fenêtre (de 3 ou 4 mots) est considérée comme suffisante.

La **catégorie grammaticale** du mot ambigu constitue pour Yarowsky (*ibid*.) un autre facteur de variation de la taille de la fenêtre : une grande fenêtre textuelle peut être utilisée pour la désambiguïsation des noms, mais pour les verbes et les adjectifs, sa taille doit être beaucoup plus petite. En effet, Gale *et al*. (1992a, 1992b) montrent que l'utilisation d'un contexte large (±50 mots autour du mot polysémique) améliore sensiblement les résultats de la désambiguïsation des noms polysémiques, par rapport à l'utilisation d'un contexte plus restreint (±6 mots).

La catégorie grammaticale du mot ambigu peut également expliquer le besoin de recourir à un **contexte symétrique**. Audibert (2003) soutient que, contrairement aux noms et aux adjectifs pour lesquels la majeure partie de l'information levant l'ambiguïté se situe au sein d'un contexte de ± 1 mot autour du mot ambigu, pour les verbes la partie essentielle de l'information se trouve en

[142] Choueka et Lusignan soulignent que pour des distinctions sémantiques plus raffinées, la validité des résultats obtenus au niveau des homographes n'est pas évidente.

position +2, voire même +3[143]. Un contexte **dissymétrique** (-2 / +4) serait donc préférable dans le cas des verbes. Crestan *et al*. (2003) ont élaboré, quant à eux, une méthode qui identifie automatiquement la fenêtre optimale pour chaque phrase contenant une instance du mot ambigu ; ce qui élimine le besoin d'identifier *a priori* la fenêtre optimale pour un mot donné. Ils ont ainsi démontré qu'utiliser un système à adaptation dynamique améliore la performance de la désambiguïsation, surtout en ce qui concerne les noms et les adjectifs.

1.3.2. Traits contextuels diversement appréhendés

Aux divergences quant à la taille du contexte viennent s'ajouter des divergences concernant la manière dont les traits du contexte doivent être pris en compte. Les informations contextuelles retenues peuvent caractériser, simplement, la présence ou l'absence de mots dans le contexte du mot ambigu, approche appelée **sac de mots**. Dans une telle approche, le contexte comprend les mots qui apparaissent à l'intérieur d'une fenêtre autour du mot ambigu, c'est-à-dire ses cooccurrents, et ceux-ci sont considérés comme un groupe en soi, sans égards à leurs relations avec le mot ambigu en termes de distance, de relations grammaticales, syntaxiques, etc. (Ide et Véronis, 1998)[144].

Les relations entretenues entre cooccurrents, ou entre cooccurrents et mot ambigu, peuvent être prises alternativement en compte (Leacock *et al*., 1998). Les informations utilisées pour la désambiguïsation sont dans ce cas plus sophistiquées et peuvent s'appliquer à **l'ordre**, la **position** et la **distance** des mots, leurs **relations syntaxiques** et **grammaticales**, les **collocations**, mais aussi à des aspects comme les **préférences de sélection** des mots, leurs **catégories sémantiques**, leurs **propriétés orthographiques** ou **morphologiques** et leur **partie du discours** (Hearst, 1991 ; Yarowsky, 1993 ; Bruce et Wiebe, 1994a, 1994b ; Leacock *et al*., 1998). De tels traits peuvent être utilisés pour filtrer les cooccurrents, dans la mesure où, généralement, la totalité des mots apparaissant

[143] Ceci s'explique par le fait que la désambiguïsation des verbes se fait davantage en fonction de leur objet que de leur sujet dans la mesure où la forme *sujet-verbe-complément* est la plus fréquente (Audibert, 2003).

[144] Cette manière de considérer le contexte se retrouve dans la plupart des approches basées sur les dictionnaires, qui ne différencient d'aucune manière le contexte et ne traitent que la cooccurrence des mots dans la même fenêtre.

dans une fenêtre ne sont pas traités. Par exemple, pour presque toutes les méthodes proposées, les informations considérées comme utiles à la désambiguïsation concernent les **mots de contenu** (en particulier, les noms, les adjectifs et les verbes) qui peuvent être repérés et sélectionnés à l'aide de l'étiquetage morphosyntaxique[145].

1.3.3. Pertinence des traits contextuels

Enfin, un autre paramètre de variation entre les différentes méthodes concerne la prise en compte de la **pertinence** des traits contextuels retenus pour la désambiguïsation. Cette pertinence peut être estimée en termes de **nature** ou de **fréquence**. La nature des mots peut, par exemple, être prise en compte en considérant les mots qui appartiennent à la **chaîne lexicale** du mot ambigu (Hirst et St-Onge, 1998 ; Vasilescu *et al*. 2004), c'est-à-dire les mots qui lui sont liés par des relations de cohésion (Halliday et Hasan, 1976). Les fréquences de cooccurrence observées peuvent être utilisées à **l'état brut**, c'est-à-dire telles qu'elles sont calculées à partir des textes, ou bien transformées par une fonction qui réduit l'effet des différences de fréquence des mots. Dans ce dernier cas, il peut s'agir d'un **test de signification**, qui compare le calcul des cooccurrences observées avec le calcul des cooccurrences attendues. A l'aide d'une telle fonction, les associations significatives entre mots sont renforcées tandis que celles qui sont accidentelles sont affaiblies. L'estimation de la pertinence des cooccurrents pour la désambiguïsation d'un mot ambigu se fait, le plus souvent, par prise en compte de leur fréquence totale dans le corpus et de leur fréquence de cooccurrence avec le mot ambigu.

Il faut noter que des divergences existent entre les différentes méthodes quant à leur façon de pondérer les traits en fonction de leur pertinence. Schütze (1998) utilise ainsi une mesure élaborée dans le domaine de la recherche d'information, la **fréquence du terme/fréquence inverse de document** (*term*

[145] L'existence d'un seul mot de contenu dans le contexte est supposée réduire l'ambiguïté de manière plus efficace qu'un contexte constitué de mots fonctionnels (articles, prépositions, conjonctions, etc.) et ce, en raison du contenu sémantique plus important de ces mots (Kaplan, 1955).

frequency/inverse document frequency - 'tf/idf') [146]. Rapp (2003) applique le **logarithme de la vraisemblance** (*log-likelihood*) (Dunning, 1993) aux calculs bruts de cooccurrence des mots pour estimer la force de leur association. En revanche, Yarowsky (1992) et Ferret (2004a) utilisent l'**information mutuelle** (Fano, 1961 : 28 ; Church et Hanks, 1990 ; Manning et Schütze, 1999 : 178-182) pour filtrer les cooccurrences les moins significatives.

1.4. Combinaison des informations du domaine et du contexte local

Quelle que soit l'approche considérée, le contexte local est en général estimé comme une source d'informations plus raffinées que celles liées au sujet traité et comme un bon indicateur de sens lors de l'utilisation d'un classificateur statistique (Leacock *et al*., 1998). Cependant, bien que les différents types d'informations contextuelles (informations du domaine et du sujet traité, informations du contexte local) soient habituellement distingués dans la littérature, l'importance de cette distinction n'est pas évidente, comme ne l'est pas non plus l'importance respective de chacun de ces types d'informations sur la sélection des sens[147]. Dans certains travaux, les informations des différents types sont même combinées (Leacock *et al*., 1996, 1998 ; Yarowsky, 1992 ; Gale *et al*., 1993). Le contexte local est censé davantage contribuer à la désambiguïsation mais les résultats sont néanmoins améliorés en étendant le contexte autour du mot ambigu. Leacock *et al*. et Yarowsky soutiennent pourtant que l'importance d'une telle combinaison dépend de la catégorie grammaticale du mot ambigu. Le bénéfice pouvant en résulter est plus important dans le cas des noms et moins évident dans le cas des verbes et des modificateurs, pour lesquels la considération du contexte local suffit généralement.

[146] Cette mesure permet d'évaluer l'importance d'un mot relativement à un document extrait d'un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document et varie en fonction de la fréquence du mot dans le corpus. La fréquence du mot est le nombre d'occurrences de ce mot dans un document. La fréquence inverse de document mesure l'importance du mot dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion des documents du corpus qui contiennent le mot.

[147] Une alternative pourrait consister à situer le micro-contexte et le contexte relatif au sujet traité dans un continuum et à considérer le rôle et l'importance des informations contextuelles comme une fonction de la distance au mot ambigu (Ide et Véronis, 1998).

1.5. Exploitation du contexte local dans un cadre bilingue

Le contexte local constitue une source importante d'informations pour des méthodes de désambiguïsation développées dans un cadre bilingue. Dans un tel cadre, l'objectif de la désambiguïsation des mots polysémiques source est de sélectionner la traduction correcte pour de nouvelles instances de ces mots. Les informations contextuelles peuvent provenir de la LS, de la LC ou des deux langues impliquées. Pourtant, dans la plupart des méthodes, ces informations proviennent uniquement du contexte de la LS.

Dans la méthode de Brown *et al*. (1991a,b), des indices trouvés dans le contexte local (± 1 mot autour du mot ambigu) sont utilisés pour construire des questions binaires permettant de choisir entre deux sens d'un mot celui qui est véhiculé par une nouvelle instance. Une fois le sens correct établi, la nouvelle instance du mot est étiquetée avec ce sens, ce qui rend possible sa traduction correcte[148].

Les méthodes de Dagan *et al*. (1991), Dagan et Itai (1994) et Golan *et al*. (1988) prennent en considération le contexte local de manière plus sophistiquée : les informations utilisées ne concernent pas la simple cooccurrence des mots avec d'autres mots au sein d'une fenêtre, mais leur **contexte syntaxique**. Ces méthodes combinent les informations syntaxiques de la LS avec des informations venant de la LC. La méthode de Golan *et al*. (*ibid*.) utilise, pour la sélection entre des traductions alternatives fournies par un dictionnaire, des règles de différenciation lexicale relatives au contexte syntaxique des mots. Plus précisément, une règle définit une correspondance entre un environnement syntaxique du mot source et une de ses traductions possibles dans la LC.

Dans la méthode de Dagan *et al*. (*ibid*.), le contexte syntaxique du mot source permet d'éliminer certaines des traductions possibles et de lever une partie de son ambiguïté. Les informations de la LC servent de filtre supplémentaire pour les traductions possibles du mot qui n'ont pas été éliminées à l'aide du contexte de la LS. Dans ce travail également, la désambiguïsation du mot source est assimilée à la tâche de sélection de l'équivalent de traduction le plus adéquat

[148] Cette méthode ne parvient à distinguer qu'entre deux sens seulement des mots ambigus, qui doivent, de plus, avoir des traductions distinctes dans la LC.

dans la LC, parmi les candidats de traduction fournis par le dictionnaire[149]. Cette assimilation des tâches se rencontre régulièrement dans les méthodes de désambiguïsation lexicale développées dans un cadre bilingue ou multilingue (Vickrey *et al.*, 2005).

1.6. Le contexte local dans notre travail

La méthode de désambiguïsation lexicale que nous proposons exploite, elle aussi, des informations provenant du contexte lexical local des mots. Il s'agit, plus précisément, d'informations de cooccurrence lexicale de premier ordre, et dont la prise en compte varie en fonction de leur pertinence pour la désambiguïsation. Il faut noter que cette étude ne tient pas compte de la structure syntaxique ni des informations positionnelles des mots dans les textes et ce, malgré le rôle important que jouent ces sources d'informations sur la désambiguïsation. Une analyse syntaxique aurait effectivement pu s'avérer utile dans l'étude de la similarité distributionnelle des mots que nous avons réalisée. De même, les informations positionnelles des mots auraient pu représenter une aide précieuse, aussi bien lors de l'étape de désambiguïsation qu'à celle de la prédiction de traduction pour de nouvelles occurrences des mots polysémiques étudiés. Dans ce travail, nous avons fait le choix de n'explorer que ce qui relève des informations de cooccurrence lexicale dans un cadre bilingue, où celles-ci sont combinées avec des informations traductionnelles. Cette démarche permet ainsi à nos méthodes d'être appliquées, entre autres, à des langues ne disposant pas d'outils d'analyse syntaxique.

Rappelons que les informations contenues dans un contexte (proche ou lointain) facilitent la sélection du sens véhiculé par une nouvelle instance d'un mot ambigu, parmi les sens possibles du mot. Pour que cette sélection soit effectuée, il est nécessaire qu'un inventaire de sens lexicaux soit disponible. De nombreuses méthodes de désambiguïsation lexicale proposées exploitent ainsi les informations fournies dans des ressources préétablies. Dans le paragraphe suivant, nous allons présenter certains aspects du fonctionnement de ces

[149] Chaque traduction possible dans la LC est considérée comme correspondant à un sens différent du mot source.

méthodes et analyser les avantages et les inconvénients qu'implique l'utilisation d'une ressource lexicale préétablie pour la désambiguïsation.

2. Désambiguïsation lexicale basée sur des connaissances

2.1. Recours à des ressources externes pour la désambiguïsation dans un cadre monolingue

2.1.1. Sources de connaissances manuellement élaborées

Les premières méthodes de désambiguïsation lexicale ont été développées dans le cadre de l'**Intelligence Artificielle** (IA), au sein de systèmes plus larges visant la compréhension de la langue naturelle. Une des approches adoptées dans ce cadre (Quillian, 1968) repose sur l'utilisation de **réseaux sémantiques** représentant des **mots** (occurrences), des **concepts** (formes) et les **relations** sémantiques les associant. La désambiguïsation s'effectue, dans ces modèles, par la recherche des **chemins d'association** les plus courts entre les nœuds activés par les mots d'entrée.

Un autre type d'approche concerne la construction de **réseaux de cadres** contenant des informations sur les mots, leurs rôles et leurs relations (Hayes, 1977 ; Hirst, 1987). Hirst a introduit les **mots polaroids**, mécanisme qui élimine progressivement les sens inappropriés des mots sur la base d'évidence syntaxique et de relations sémantiques observées dans le réseau[150].

Small (1979) a proposé l'utilisation de **systèmes experts** pour la désambiguïsation[151]. Dans une telle approche, un système expert est créé pour chaque mot qui contient un **réseau de discrimination** des sens de mot. Ce réseau est parcouru sur la base d'informations fournies par le contexte et par d'autres experts ; à la fin, un seul sens demeure, qui est ensuite ajouté à la représentation sémantique de la phrase.

[150] Pourtant, lorsqu'un mot est employé au sein d'une phrase dans un sens métaphorique, métonymique ou non connu, les polaroids finissent souvent par éliminer tous les sens possibles et échouent. Si cette approche est efficace pour la désambiguïsation des homonymes, elle ne l'est pas pour d'autres types de polysémie.

[151] L'inconvénient principal de cette approche est que les experts doivent être extrêmement larges et complexes pour parvenir à ce but.

Autre approche, très importante, celle de la **sémantique préférentielle** proposée par Wilks (1975 ; Wilks et Fass, 1992), qui spécifie des restrictions de sélection concernant les combinaisons d'éléments lexicaux dans une phrase, à l'aide de traits sémantiques. Cette approche vise à attribuer l'interprétation la plus « cohérente » à une phrase, en termes de satisfaction du nombre maximal possible de préférences internes de ses parties.

Les méthodes de désambiguïsation développées dans le cadre de l'IA sont des **méthodes** essentiellement **connexionnistes**. Ces méthodes sont basées sur le principe d'**amorçage sémantique**, d'après lequel l'introduction d'un concept influence et facilite le traitement des concepts introduits par la suite et qui lui sont sémantiquement liés. Cette idée a été implémentée dans les modèles de **propagation de l'activation** (Collins et Loftus, 1975 ; Quillian, *ibid*. ; Cottrell et Small, 1983 ; Waltz et Pollack, 1985), où les concepts d'un réseau sémantique sont activés lors de leur utilisation et où l'activation se propage aux nœuds connectés. Outre les liens d'activation, ces réseaux peuvent aussi contenir des liens d'**inhibition** qui servent à supprimer, certains voisins du nœud activé.

La difficulté et le coût de l'encodage manuel des sources de connaissance nécessaires aux systèmes développés dans le cadre de l'IA les ont restreints à des implémentations qui traitent de minuscules parties de la langue (*toy implementations*) et ont empêché la généralisation du travail effectué en dehors de ces domaines très limités. En outre, l'évaluation de ces procédures de désambiguïsation était réalisée sur de petits ensembles de test et dans un contexte limité, ce qui rendait difficile la détermination de leur efficacité sur des textes réels. L'apparition de ressources lexicales informatisées de grande envergure dans les années 80 a ouvert la voie au développement de méthodes de désambiguïsation capables d'exploiter les informations fournies dans ces ressources, et permettant d'éviter l'étape, longue et fastidieuse, de l'encodage manuel.

2.1.2. Ressources lexico-sémantiques informatisées

Les dictionnaires et les autres inventaires sémantiques ont donc constitué une source alternative d'informations exploitables pour la désambiguïsation

lexicale. Lorsqu'ils sont disponibles sur support électronique, ces ressources sont alors directement exploitables par les méthodes automatiques de désambiguïsation[152]. Les méthodes qui exploitent des informations d'une ressource externe sont caractérisées comme des méthodes **dirigées par les connaissances**[153].

La première méthode de ce type à être proposée, et qui a eu une grande influence sur les méthodes qui ont suivi, est celle de Lesk (1986). Dans le cadre de cette méthode, la sélection du sens véhiculé par une nouvelle instance d'un mot ambigu se fait en calculant le **recouvrement** entre les mots inclus dans les définitions des sens du mot et ceux inclus dans les définitions des cooccurrents de sa nouvelle instance au sein d'un dictionnaire informatisé[154]. Le sens sélectionné est celui dont la définition contient le plus grand nombre de mots communs avec les définitions des sens des mots du nouveau contexte. La désambiguïsation a donc lieu en choisissant, pour le mot ambigu et les mots qui l'entourent, les définitions qui se recoupent le plus[155].

Le principal inconvénient de cette méthode est de reposer sur la **correspondance exacte** entre les mots trouvés dans les définitions dictionnairiques. Cette exigence de correspondance exacte la fait donc dépendre fortement des mots utilisés dans les définitions et la rend très sensible à la présence (ou non) d'un mot au sein de ces définitions. Elle ne lui permet pas, en outre, de capter des relations moins directes entre les mots, c'est-à-dire des relations qui ne sont pas explicitement décrites dans les définitions (Véronis et Ide, 1990). Malgré cet inconvénient, l'idée principale de la méthode de Lesk a été reprise et élaborée dans de nombreux travaux qui ont suivi.

[152] Ressources lexicales fréquemment utilisées : dictionnaires unilingues de langue générale, comme le 'COLLINS COBUILD English Dictionary', le 'LDOCE' ('Longman Dictionary of Contemporary English'), l''Oxford English Dictionary' ('OED'), l''Oxford Advanced Learners Dictionary', le 'Merriam-Webster' et l''American Heritage Dictionary of the English Language' pour l'anglais, et les 'Petit Robert' et 'Petit Larousse' pour le français ; dictionnaires bilingues, comme les dictionnaires anglais-français 'Robert & Collins' et 'Oxford-Hachette' et des thésaurus, comme le 'Roget's Thesaurus' pour l'anglais et le thésaurus 'Larousse' pour le français.

[153] A différencier des méthodes « dirigées par les données », décrites dans le paragraphe suivant.

[154] Dans cette méthode, les cooccurrents sont les mots qui apparaissent dans une fenêtre textuelle de dix mots autour de la nouvelle instance du mot ambigu.

[155] Des variantes de l'approche de Lesk ont été utilisées en tant qu'approches « baseline » dans Senseval 1 et 2 (Kilgarriff et Rosenzweig, 2000a, b ; Kilgarriff, 2001 ; Edmonds, 2002).

La méthode proposée par Wilks *et al*. (1990) permet d'estimer la similarité entre entrées de sens[156] et contextes, même s'ils ne partagent pas de mots en commun. Cette manière de procéder est rendue possible par l'**expansion** des entrées de sens du dictionnaire et des contextes à l'aide de **données de cooccurrence**, collectionnées à partir des définitions des sens[157]. Cette expansion se fait par l'inclusion de mots liés aux mots présents dans les contextes et les entrées de sens. Un vecteur de mots est alors construit pour l'entrée de chaque sens et un autre pour le contexte (la phrase où le mot apparaît), en ajoutant les vecteurs des mots liés à chacun des mots de l'entrée ou du contexte, respectivement, et en excluant le mot ambigu. Le sens retenu est celui dont le vecteur est le plus similaire au vecteur du contexte.

Cette idée de désambiguïsation par mise en évidence des liens sémantiques entre les mots utilisés dans une phrase existe déjà dans la méthode proposée par Sparck Jones (1986) [158], qui vise l'identification du sens des mots dans un texte et le repérage du sujet traité. Cette méthode consiste à représenter chaque mot d'un texte par une liste des entrées dans lesquelles il apparaît dans un thésaurus (1986 : 22-23)[159]. Etant donné le nombre élevé de mots ambigus, chaque mot est représenté par une liste contenant plus d'une entrée du thésaurus. La comparaison des listes attribuées aux mots différents d'un texte permet de repérer les entrées qui apparaissent dans plus d'une liste ; celles-ci spécifient l'usage du mot ambigu et indiquent le sujet traité par le texte.

Une des différences distinguant la méthode de Sparck Jones de celle de Wilks *et al*. (*ibid*.) est que cette dernière n'est pas capable de désambiguïser tous les mots de la phrase en même temps, à cause du phénomène d'**explosion combinatoire**. Le fait que la désambiguïsation simultanée de plus d'un mot ambigu soit impossible constitue, pour Véronis et Ide (1990), la faiblesse

[156] L'entrée d'un sens comprend la définition du sens et un exemple d'utilisation du sens.
[157] La fréquence de cooccurrence de deux mots correspond au nombre d'entrées de sens dans lesquelles les deux mots apparaissent.
[158] Il s'agit de la publication de la thèse de doctorat de Spack Jones, soutenue en 1964.
[159] Le but est de remplacer les mots du texte par des notions plus générales. Les étiquettes de sujet étant considérées comme très grossières et limitées, Sparck Jones préfère utiliser une classification sémantique adéquate pour tous les mots du vocabulaire d'une langue, qui traite tous les usages des mots et pas uniquement ceux pour lesquels des étiquettes du domaine peuvent être utilisées. Ceci justifie la décision de se référer à un thésaurus ('Roget's'), où les mots sont classifiés relativement aux idées qu'ils expriment.

principale de la méthode de Wilks *et al*. Celle-ci se heurte, en outre, aux problèmes déjà signalés pour la méthode de Lesk. La réponse de Véronis et Ide (*ibid*.) consiste à construire de grands **réseaux de neurones** dédiés à la désambiguïsation. Les nœuds de ces réseaux représentent des mots et sont connectés par des liens d'**activation** et, éventuellement, des liens d'**inhibition** qui connectent des sens antagonistes d'un mot. Les nœuds correspondant aux mots de la phrase analysée sont d'abord activés et ces mots activent ensuite leurs voisins qui activent à leur tour leurs propres voisins. Le réseau se stabilise progressivement dans un état où un sens de chaque mot d'entrée est plus activé que les autres. Pour la construction automatique de réseaux de ce type, Véronis et Ide exploitent également les informations contenues dans des définitions dictionnairiques, en se fondant sur l'hypothèse de relations sémantiques pertinentes entre un mot et les mots utilisés pour le définir. Les connexions du réseau reflètent ces relations. Cette méthode améliore les résultats obtenus par les méthodes précédentes sur des cas précis, tout en n'ayant pas besoin d'encodage d'informations sémantiques.

D'autres améliorations de la méthode de Lesk ont été également proposées dans des travaux ultérieurs. Guthrie *et al*. (1991) exploitent les classifications de sujet fournies dans le 'LDOCE' pour établir des liens de cooccurrence, dépendants du sujet, entre les mots utilisés dans les définitions. Les voisinages lexicaux construits de cette manière sont ainsi dépendants du domaine et la désambiguïsation des nouvelles instances d'un mot ambigu s'opère en calculant le recouvrement entre les voisinages créés pour chacun de ses sens et les mots du contexte. La méthode de Cowie *et al*. (1992), quant à elle, enrichit les définitions dictionnairiques du 'LDOCE' avec des informations de domaine, en traitant le code de domaine attribué à un sens comme un mot faisant partie de sa définition. La sélection correcte des sens des mots repose sur l'idée que les sens lexicaux qui apparaissent dans la même phrase ont plus de mots et de codes de sujet en commun dans leurs définitions[160] que ceux appartenant à des phrases

[160] Cette idée est analysée et défendue par Wilks *et al*. (1990), qui soutiennent que : a) la probabilité d'une relation entre deux sens lexicaux qui apparaissent dans la même phrase est suffisamment élevée pour rendre possible l'extraction d'informations utiles à partir de statistiques de cooccurrence ; b) le degré auquel cette probabilité dépasse la probabilité de cooccurrence imputable au hasard constitue un indicateur de la force de la relation et c) si lors d'une attribution de sens aux

différentes[161]. En revanche, la désambiguïsation par la méthode de Krovetz et Croft (1992) s'effectue en déterminant le code de domaine (issu de 'LDOCE') qui reçoit le score le plus élevé dans une fenêtre contextuelle. Ce code est ensuite utilisé pour augmenter le poids des sens auxquels il est attribué[162].

Yarowsky (1992) désambiguïse les mots d'un texte en se servant des distinctions sémantiques représentées dans les catégories du thésaurus 'Roget's', considérées comme des approximations de classes conceptuelles. La méthode se base sur l'observation de l'apparition de classes conceptuelles différentes au sein des contextes[163], et de la tendance des sens d'un mot à appartenir à des classes différentes. Un discriminateur contextuel élaboré pour les classes conceptuelles est utilisé en tant que discriminateur des sens lexicaux appartenant à ces classes. Par conséquent, les indicateurs de contexte d'une catégorie du thésaurus sont considérés comme des indicateurs de contexte pour les membres de cette catégorie.

Malgré la large exploitation des informations sémantiques contenues dans les dictionnaires informatisés par les méthodes de désambiguïsation, leur qualité est souvent remise en question. Ces dictionnaires, développés pour un usage humain, ne sont généralement pas dotés de la systématicité et de la finesse de description requises dans un cadre automatique. Ils présentent, en outre, un fort degré de **divergence** au niveau de la représentation des sens lexicaux et des relations entretenues entre eux. Des critiques ont également été faites contre l'utilisation de thésaurus informatisés pour la désambiguïsation. Bien que ces ressources fournissent une catégorisation sémantique et des réseaux riches d'associations entre mots, les niveaux supérieurs des hiérarchies conceptuelles

mots d'une phrase, le nombre et la force des relations entre les sens lexicaux sont plus grands que lors d'une autre attribution, la première a davantage de chances d'être correcte.

[161] La méthode utilisée (« recuit simulé », *simulated annealing* en anglais) permet la détermination simultanée de tous les sens des mots dans une phrase. Le nombre de sens de chaque mot correspond à celui trouvé dans le 'LDOCE'.

[162] La méthode ne vise pas nécessairement l'identification d'un seul sens correct pour un mot, mais plutôt l'élimination du plus grand nombre possible de sens incorrects et l'attribution d'un poids élevé aux sens probablement corrects.

[163] Des mots indicatifs de chaque catégorie du thésaurus sont repérés en collectionnant des contextes représentatifs de la catégorie, constitués d'ensembles de cooccurrents des membres de la catégorie dans les textes. La matrice pondérée construite sert d'exemple de contexte typique de la catégorie. La présence des mots significatifs d'une catégorie dans le contexte d'un mot ambigu révèle l'évidence de son appartenance à la catégorie indiquée.

sont souvent discutables et caractérisés comme trop larges pour être utiles à l'établissement de catégories sémantiques significatives.

2.1.3. Ressources lexico-sémantiques électroniques

La mise en évidence des faiblesses des ressources informatisées pour le traitement automatique a généré l'émergence de bases de connaissances de grande envergure élaborées manuellement, comme les **dictionnaires électroniques**. L'exemple le plus connu, qui est aussi le dictionnaire électronique le plus largement utilisé dans un cadre automatique, est le réseau sémantique **WordNet** (Miller *et al.*, 1990). Cette ressource pourrait être caractérisée comme un **lexique d'énumération** en raison de la représentation explicite des sens lexicaux. Comme nous l'avons déjà dit, les sens sont représentés, dans WordNet, à l'aide de **synsets**, ensembles de mots synonymes représentant un concept lexical. Les synsets sont organisés au sein d'une **hiérarchie conceptuelle** et sont liés par des relations sémantiques, comme l'hyponymie, l'hypéronymie, l'antonymie et la méronymie.

Certaines méthodes de désambiguïsation exploitant WordNet (Voorhees, 1993 ; Sussna, 1993 ; Richardson et Smeaton, 1995 ; Resnik, 1995) profitent des informations taxonomiques incluses et utilisent des métriques qui calculent la **distance** (ou la similarité) **sémantique** entre les mots d'entrée, pour les désambiguïser. Certaines de ces méthodes, comme celle de Sussna, calculent la distance entre les mots à l'aide des **arêtes** qui lient les synsets correspondants de WordNet. Cette méthode repose sur l'idée que les sens corrects d'un ensemble de mots apparaissant à proximité dans un texte sont ceux décrits par les synsets qui minimisent la distance entre les mots en question au sein du réseau. D'autres méthodes, comme celle de Resnik, calculent le **contenu informationnel** commun aux mots sur la base de l'hypothèse que les sens corrects à attribuer aux mots polysémiques apparaissant ensemble sont ceux qui partagent des éléments de sens.

Un autre type de méthodes basées sur WordNet combine les informations taxonomiques à celles trouvées au sein des définitions de sens lexicaux. Tel est le

cas de la méthode de Banerjee et Pedersen (2002), fondée sur l'approche de désambiguïsation de Lesk. Mais au lieu d'utiliser les définitions de dictionnaires traditionnels, leur méthode exploite les informations contenues dans les relations lexicales définies par WordNet. L'algorithme de Lesk repose sur la révélation de recouvrements entre les définitions dictionnairiques de mots voisins au mot à désambiguïser, tandis que celui de Banerjee et Pedersen étend les comparaisons aux définitions de mots liés à la fois au mot ambigu et aux mots de son contexte, au sein de WordNet. La richesse des informations ainsi exploitées améliore la précision de la désambiguïsation.

Vasilescu *et al*. (2004) analysent de façon détaillée les paramètres déterminant la performance des méthodes de désambiguïsation basées sur l'algorithme de Lesk et exploitant les informations de WordNet. Cette analyse s'effectue en comparant les variantes de l'algorithme, variantes relatives à la manière dont le contexte des mots ambigus est considéré[164], la manière dont les sens sont décrits[165], leur pondération[166] ainsi que les mots du contexte pris en compte[167]. Patwardan *et al*. (2003) prolongent l'étude de Banerjee et Pedersen (*ibid*.) : en considérant le recouvrement des définitions comme une mesure de similarité sémantique, ils procèdent à la désambiguïsation en utilisant d'autres mesures de similarité sémantique sur la base des informations de WordNet. Naskar et Bandyopadhyay (2007), en revanche, utilisent l'algorithme de Lesk

[164] Il s'agit soit de prendre en compte les descriptions dans WordNet de tous les sens des mots qui apparaissent dans le contexte soit, plus simplement, de ne considérer que les mots du contexte en ignorant leurs sens. La deuxième variante a donné de meilleurs résultats.

[165] Un sens peut être décrit par l'ensemble des lemmes de mots pleins associés à la définition du sens fournie dans le champ correspondant de WordNet, ou associés aux exemples. Une alternative consiste à prendre en compte les synonymes (synset) du sens et tous les synsets qui entretiennent une relation d'hyperonymie avec lui, jusqu'au sommet de la hiérarchie WordNet. Ces deux approches peuvent se combiner.

[166] La pondération simple définit l'attribution d'un score au sens candidat qui correspond au nombre de recouvrements entre les informations associées au sens (mots dans sa définition, etc.) et les informations associées au contexte. La pondération peut aussi être effectuée en prenant en compte la longueur de la description du sens, dans la mesure où des descriptions plus longues peuvent produire davantage de recouvrements que des descriptions plus courtes ; ces descriptions dominent ainsi le processus de prise de décision (Lesk, 1986). Mais le fait que ce paramètre influence beaucoup le résultat de la désambiguïsation n'a pas été démontré.

[167] Cela concerne tous les mots pleins du contexte ou seulement les mots appartenant à la « chaîne lexicale » du mot ambigu (Hirst et St-Onge, 1998), qui est identifiée sur la base des relations de synonymie et d'hypéronymie des mots dans WordNet. Un mot appartient à la chaîne lexicale du mot ambigu si les ensembles de synonymes et d'hyperonymes des sens de ces mots présentent une similarité assez forte. La considération de la chaîne lexicale des mots ambigus a amélioré la performance de la désambiguïsation.

dans un système de désambiguïsation basé sur 'Extended WordNet' (Harabagiu *et al*., 1999), où les définitions des synsets sont étiquetées par des informations sémantiques et morphosyntaxiques.

Toutes les méthodes de désambiguïsation qui exploitent les ressources prédéfinies décrites jusqu'ici ont été développées dans un cadre monolingue. Des ressources lexicales préétablies ont pourtant été utilisées par les méthodes développées dans un cadre bi- et multi-lingue. L'objectif de la désambiguïsation lexicale menée dans un tel cadre est, le plus souvent, de sélectionner l'équivalent de traduction correct, dans une LC, pour de nouvelles occurrences des mots polysémiques d'une LS. Le fonctionnement de ces méthodes fera l'objet du paragraphe suivant.

2.2. Recours à des ressources externes pour la désambiguïsation dans un cadre bi-(multi-) lingue

2.2.1. Resources lexico-sémantiques informatisées

Brun *et al*. (2001) utilisent un dictionnaire électronique bilingue français-anglais (le 'Oxford-Hachette French Dictionary') en tant que **corpus sémantiquement étiqueté**, dans le but d'en extraire un ensemble de règles de désambiguïsation sémantique. Ce dictionnaire contient de nombreux exemples représentatifs de chaque sens des mots français, ainsi que des informations riches sur les collocations typiques qui pointent sur des équivalents de traduction différents. Les règles extraites à partir du dictionnaire sont de deux types : des **règles lexicales**, qui s'appliquent sur la base du contexte lexical du mot à désambiguïser, et des **règles sémantiques** construites en utilisant les catégories sémantiques d'un dictionnaire ('AlethDic'), qui s'appliquent sur la base du contexte sémantique du mot à désambiguïser. Les règles lexicales de désambiguïsation sont élaborées à partir du texte illustrant chaque sens lexical (exemples d'utilisation, collocations, etc.), à l'aide d'un analyseur syntaxique de surface, qui en extrait les relations fonctionnelles (relations prédicats-arguments) et syntaxiques (de type sujet, objet, modifieur, etc.). Les classes sémantiques, quant à elles, permettent d'élargir le champ d'application des règles lexicales, en

remplaçant les arguments des relations fonctionnelles par l'ensemble des classes sémantiques auxquelles ils appartiennent.

La base de ces règles est utilisable pour la désambiguïsation de nouvelles instances des mots ambigus. Pendant la phase de désambiguïsation, la phrase contenant le mot ambigu est syntaxiquement analysée et les relations fonctionnelles mettant en jeu le mot sont extraites et comparées aux règles de désambiguïsation. Si une règle s'applique, le numéro du sens correspondant dans le dictionnaire est attribué au mot et la traduction proposée est celle qui correspond à l'exemple (ou à la collocation) à partir duquel la règle lexicale qui s'applique avait été construite.

Dufour (1997) et Segond *et al*. (2000) proposent un ensemble de mécanismes de désambiguïsation opératoires dans le cadre de systèmes de compréhension multilingue. Le système de Dufour (1997) cherche des correspondances entre le contexte des nouvelles instances de mots ambigus et les informations linguistiques et métalinguistiques (partie du discours, restrictions de collocations, étiquettes de domaine et de style) extraites de dictionnaires bilingues[168], afin de choisir la traduction correcte. En revanche, Segond *et al*. extraient des informations de sous-catégorisation et de collocation à partir d'un dictionnaire bilingue[169], qui concernent le type de sujet et/ou d'objet attendu par un prédicat. Des relations de dépendance (sujet-objet) sont, par la suite, extraites des nouveaux textes à l'aide d'un analyseur syntaxique de surface. Les relations extraites des textes sont mises en correspondance avec les informations de collocation retenues du dictionnaire et, si des correspondances sont trouvées, la traduction sélectionnée est alors proposée.

Les travaux présentés jusqu'ici exploitent les informations contenues dans des ressources informatisées. L'exploitation de ressources électroniques conformes au traitement bi- et multi-lingue a été aussi proposée.

[168] Les dictionnaires utilisés sont le 'Collins & Robert' et l''Oxford University Press-Hachette French-English, English-French Dictionary' ('OUP-H').

[169] Le dictionnaire utilisé est également le 'OUP-H'. Ce dictionnaire contient des étiquettes de collocation, qui encodent les types de sujet et/ou d'objet attendus par un prédicat.

2.2.2. Ressources lexico-sémantiques électroniques

Tufiş *et al*. (2004c) adoptent, comme nous l'avons déjà vu, une conception traductionnelle du contexte. Leur méthode de désambiguïsation présuppose l'utilisation de **corpus parallèles**. En effet, les informations contextuelles utilisées par cette méthode de désambiguïsation concernent les correspondances de traduction des mots au sein d'un corpus parallèle et sont combinées avec des informations provenant du réseau sémantique **BalkaNet**[170]. Les wordnets inclus dans BalkaNet sont alignés au Princeton WordNet, considéré comme un **index inter-langue** (*Interlingual Index* (ILI)). L'hypothèse sous-jacente à cet alignement est que des traductions réciproques dans des textes parallèles devraient avoir les mêmes sens inter-langues (ou proches) ou, en reprenant les termes de BalkaNet, les mêmes codes ILI. Ainsi, après l'étape d'alignement lexical du bitexte utilisé, qui permet l'identification de paires de mots constituant des traductions réciproques, les codes ILI des synsets contenant les traductions sont repérés. Le code ILI qui se trouve à l'intersection des deux listes de codes correspond au sens commun des mots alignés. Si l'intersection est vide, la paire de codes les plus similaires, parmi les paires candidates, est identifiée, tandis que si elle contient plus d'un code ILI, il s'agit d'un cas d'ambiguïté cross-linguistique[171].

Les méthodes de désambiguïsation que nous venons de présenter reposent sur des connaissances extraites à partir de ressources lexicales monolingues et bilingues. Nous avons analysé les différentes manières dont les informations trouvées au sein des ressources peuvent être exploitées et combinées avec des informations contextuelles pour la levée de l'ambiguïté de nouvelles instances des mots ambigus. La possibilité d'exploiter ces ressources constitue certainement une force de ces méthodes, qui s'accompagne pourtant d'un nombre d'inconvénients, dont nous avons brièvement fait mention. Les faiblesses

[170] BalkaNet comprend un ensemble de wordnets construits pour les langues suivantes : grec, bulgare, roumain, turque, serbe et tchèque (extension du wordnet tchèque développé dans le cadre d'EuroWordNet).

[171] Dans ce travail, le mécanisme de clustering présenté par Ide *et al*. (2001 ; 2002) est utilisé comme une méthode complémentaire, dans les cas où la méthode de désambiguïsation basée sur le WordNet ne parvient pas à désambiguïser une occurrence du mot ambigu, et dans les cas d'ambiguïté traductionnelle.

des méthodes qui font référence à des ressources préétablies seront analysées plus en détail dans le paragraphe suivant.

2.3. Avantages et inconvénients liés à l'exploitation de ressources externes pour la désambiguïsation

2.3.1. Disponibilité des ressources

La disponibilité de ressources de grande taille sous format électronique a été considérée comme une solution au besoin de construction manuelle de grandes bases de connaissances pour la désambiguïsation lexicale. Outre la quantité d'informations incluses dans un dictionnaire, la qualité de ces informations a également constitué un argument favorisant l'utilisation de ce type de ressources pour la désambiguïsation : un dictionnaire est en effet conçu et organisé autour des notions de sens et de mot et constitue ainsi un outil parfaitement adapté pour effectuer une sélection du sens contextuel d'un mot (Brun *et al*., 2001). Le découpage rigoureux des entrées polysémiques en sens implique, en outre, une grande précision de ces informations, ce qui désigne le dictionnaire comme une ressource d'informations sémantiques particulièrement fiable. C'est pour ces raisons qu'un grand nombre de travaux a été consacré au développement d'algorithmes permettant l'extraction d'informations fournies par ces ressources, leur modélisation et leur exploitation dans le cadre de la désambiguïsation.

Les inconvénients liés à l'utilisation de ce type de ressources dans un cadre automatique sont néanmoins régulièrement signalés dans la littérature. L'inconvénient principal concerne la disponibilité, dans des langues différentes, de ressources informatisées ou électroniques à large couverture et de bonne qualité. La plupart des méthodes de désambiguïsation basées sur les connaissances exploitent un nombre limité de ressources, surtout en anglais. L'absence de ressources correspondantes pour un grand nombre de langues constitue un obstacle important à la généralisation de l'application des méthodes basées sur ce type de connaissances.

2.3.2. Divergences qualitatives et structurales entre ressources

Les autres problèmes rencontrés lors de l'utilisation de ressources lexicales préétablies dans un cadre automatique sont liés à la quantité et à la nature des informations qu'elles fournissent (Slator et Wilks, 1987 ; Ide et Véronis, 1998). Les divergences au niveau de la nature, du nombre et de la granularité des distinctions sémantiques entre les différentes ressources sont remarquables. Nous avons déjà expliqué (cf. §1.3, chapitre 1) certaines des raisons (linguistiques et autres) qui provoquent ces divergences et qui montrent qu'un consensus à propos de la nature du sens est difficile à obtenir. Un problème pratique résulte de cette diversité et concerne la difficulté de combiner des informations provenant de différentes ressources. Or le besoin de lier différentes ressources apparaît dès que la couverture d'une seule ressource ne s'avère pas suffisante pour une tâche précise[172] ou dans le cadre d'applications où l'association d'informations de nature différente est souhaitée. Tel est le cas, par exemple, dans les travaux de Segond *et al*. (2000) et de Brun *et al*. (2001), où des informations sémantiques issues d'une ressource monolingue (respectivement le 'Petit Larousse' et le thésaurus 'AlethDic') sont combinées à des informations de traduction provenant d'un dictionnaire bilingue (le 'OUP-H', dans les deux cas).

Un paramètre qui entre en jeu, lors des tentatives de combinaison d'informations trouvées dans des ressources différentes, concerne la manière dont sont construites les ressources. Segond *et al*. (*ibid*.) soulignent la difficulté de combiner les informations provenant d'un dictionnaire monolingue traditionnel à celles fournies par un dictionnaire bilingue basé sur des corpus[173].

La combinaison de ressources monolingues et bilingues se heurte par ailleurs à leurs particularités structurales : dans les dictionnaires bilingues, le

[172] Ide et Véronis (1993) montrent comment la combinaison d'informations taxonomiques extraites automatiquement de dictionnaires différents améliore de manière significative la complétude et la précision des arbres taxonomiques qui en résultent.

[173] Le 'Petit Larousse', dictionnaire monolingue conçu pour des locuteurs français natifs, fournit une hiérarchie de sens sophistiquée. En revanche, le 'OUP-H', dictionnaire bilingue destiné à être utilisé par des locuteurs non natifs, fournit un ensemble plat de sens. En outre, le 'Petit Larousse' donne la priorité à la sémantique et ne dispense des informations syntaxiques qu'à titre indicatif, tandis que l' 'OUP-H' fournit, de manière explicite, toutes les constructions syntaxiques fréquentes et distingue un sens pour chacune d'elles. Par conséquent, lors de la mise en correspondance des distinctions sémantiques de ces deux dictionnaires, un sens de OUP-H peut correspondre à plusieurs sens du Petit Larousse ou ne correspondre à aucun.

découpage en sens est plus systématique, du fait des impératifs de correspondance avec l'autre langue. Ainsi, les sens proposés sont de granularité plus fine que ceux présents dans un dictionnaire monolingue, et le nombre d'exemples significatifs est plus important (Brun *et al.*, 2001). Le type de ressources (dictionnaire ou thésaurus) influence aussi de manière importante la possibilité de leur association car des problèmes liés à la cohérence des informations peuvent apparaître.

2.3.3. Granularité des descriptions sémantiques

2.3.3.1. Finesse des descriptions et accord entre annotateurs

Un autre paramètre qui empêche l'utilisation efficace de ressources préétablies lors des tâches de désambiguïsation et d'étiquetage sémantique concerne la granularité des sens décrits (Ng *et al.*, 1999 ; Ide *et al.*, 2001). Le nombre de sens proposés est parfois tellement important et la distance inter-sens tellement faible, qu'une distinction entre eux s'avère difficile, même par des humains[174]. En effet, le fort taux de désaccord souvent observé entre annotateurs, lors de tâches d'étiquetage par référence à des ressources sémantiques prédéfinies (Kilgarriff, 1993 ; Véronis 1998) [175], peut être imputé à la difficulté de distinguer les sens ; ce taux augmente avec la finesse de la granularité des distinctions sémantiques[176].

Une granularité élevée des sens rend possible l'attribution de différents sens à une nouvelle instance d'un mot, sans qu'aucun ne soit vraiment erroné. La difficulté à choisir entre étiquettes sémantiques raffinées augmente encore lorsque le nombre d'exemples de phrases illustrant l'usage des différents sens est

[174] Le problème posé par la granularité fine des sens a amené à distinguer deux sous-tâches, dans SemEval-2007, concernant la désambiguïsation lexicale en anglais : l'une reposant sur des sens lexicaux de granularité fine et l'autre de granularité grossière.

[175] Véronis souligne, justement, que dans une tâche d'annotation sémantique effectuée par six annotateurs, le taux d'accord entre eux est très faible et même, pour certains mots, guère plus élevé que le hasard. Les annotateurs devaient choisir, pour chaque nouvelle occurrence des mots, un ou plusieurs sens donnés par le dictionnaire (Petit Larousse) pour les mots en question.

[176] Le taux de désaccord ne diminue pas, même lorsque les annotateurs sont des lexicographes professionnels (Véronis, 1998).

faible. En revanche, dans les cas de distinctions de granularité plus grossière, qui englobent des distinctions difficilement repérables – diminuant ainsi le nombre de possibilités – l'accord entre les annotateurs augmente. Le lien clair qui s'établit entre granularité des sens et accord des annotateurs, lors des tâches d'étiquetage, désigne cet accord comme un indice assez fiable pour proposer l'identification de sens de granularité grossière (Ng *et al*., 1999 ; Véronis, 1998 ; Bruce et Wiebe, 1998)[177].

Les difficultés auxquelles se confrontent les annotateurs lors de tâches de désambiguïsation et d'étiquetage impliquant des distinctions sémantiques très fines démontrent qu'un locuteur humain, en situation réelle, n'a pas besoin d'effectuer de telles distinctions afin de comprendre et d'utiliser sa langue (Ng *et al*., 1999). Très intéressant est le travail de Jorgensen (1990) sur ce sujet, mené dans un cadre psycholinguistique et mettant en œuvre des expériences visant à estimer la réalité psychologique des sens lexicaux. Partant de l'hypothèse qu'il n'y a pas de manière précise de distinguer les sens, Jorgensen considère qu'ils peuvent être identifiés en collectionnant des jugements de locuteurs d'une langue sur la similarité sémantique et les différences d'usage des mots. Les expériences menées impliquent le regroupement d'un ensemble d'instances de mots en fonction de leur similarité sémantique[178], d'une part, et, d'autre part, relativement aux définitions de sens issues d'un dictionnaire.

Les résultats obtenus par des sujets différents lors de la première étape sont ensuite comparés, ainsi que les résultats fournis par le même sujet à l'issue des deux étapes, c'est-à-dire avant et après la présentation des définitions dictionnairiques. Les estimations du nombre de sens avant la présentation des définitions dictionnairiques (environ trois par mot), se modifient avec la deuxième étape. Le nombre des catégories sémantiques augmente de manière significative pour les mots fortement polysémiques, sans pour autant coïncider avec le nombre, encore plus élevé, de catégories proposées par le dictionnaire. La

[177] Pour Bruce et Wiebe (1998), le désaccord systématique des annotateurs concernant un ensemble de sens précis peut servir d'indice à leur « fusionnement ». L'hypothèse gouvernant ce regroupement est que les sens sont tellement proches que leurs différences ne sont pas évidentes aux annotateurs. Pour Véronis (1998), le désaccord systématique entre annotateurs peut servir d'indice au fusionnement en « super-tags », qui engloberaient les étiquettes sémantiques correspondant aux sens proches de granularité fine.

[178] Ce principe est également sous-jacent à la classification d'instances de mots effectuée par les lexicographes pour la distinction des sens lexicaux qu'ils vont ensuite décrire au sein de définitions.

principale conclusion de ce résultat est que les catégories du dictionnaire sont plus nombreuses que celles normalement distinguées par des locuteurs. En revanche, les modifications quant aux estimations concernant les mots ayant un nombre de sens définis plus petit dans le dictionnaire sont beaucoup moins remarquables.

2.3.3.2. Inconvénients liés à l'utilisation de descriptions fines pour la désambiguïsation

Le problème d'une granularité très fine des distinctions sémantiques constitue la critique principale à l'utilisation du réseau sémantique WordNet dans des applications automatiques (Kilgarriff, 1997b, 2001 ; Edmonds et Kilgarriff, 2002). Le grand nombre de sens proposés et leur proximité rendent difficile la sélection d'un sens lors des tâches de désambiguïsation. La pertinence même d'une telle sélection peut être remise en question : une approche de **choix forcé** d'un seul sens adéquat à partir d'un ensemble de sens préétabli peut conduire à des décisions arbitraires (Dolan, 1994). Retenir un sens « correct » peut provoquer une perte d'informations éventuellement utiles sur la sémantique d'un mot, surtout lorsque les liens possibles entre les sens ne sont pas pris en compte[179]. La confrontation d'algorithmes de désambiguïsation à de multiples choix « corrects » complique sensiblement la sélection d'un sens raisonnable, tandis que la discrimination entre sens très fortement similaires peut provoquer un gaspillage de ressources sans bénéfice clair. La combinaison d'un lexique encodant des distinctions sémantiques inutilement fines à un algorithme de désambiguïsation qui entreprend la tâche artificielle de ne retenir qu'un seul sens correct du mot, implique que la quantité d'informations sémantiques extraites pour un mot soit toujours limitée à ce qui est disponible dans un sens individuel,

[179] Ce phénomène constitue une différence importante entre le traitement de ce type d'informations par les utilisateurs humains et le traitement automatique. Les utilisateurs humains, au lieu de traiter les sens décrits pour un mot dans le dictionnaire comme complètement distincts, parviennent à une notion plus abstraite du sens du mot, qui peut combiner des informations de sens distincts. Pour qu'un système automatique parvienne au même résultat, il faut que la manière dont les sens lexicaux se recouvrent soit encodée explicitement.

tandis que d'autres informations importantes à propos de sa sémantique sont éventuellement perdues.

Hormis la complexité générée, lors du traitement, par la granularité élevée des distinctions sémantiques et la perte d'informations due à la sélection forcée d'un seul sens, qui constituent déjà des problèmes importants, la nécessité d'opérer des distinctions de ce type est aussi remise en cause. Des applications comme l'Inférence Textuelle et le Traitement de Connaissances, par exemple, nécessitent de disposer d'un grand ensemble de relations entre concepts (Mihalcea et Moldovan, 2001 ; Harabagiu et Moldovan, 1998). La multitude de concepts et de liens sémantiques existant dans WordNet explique qu'il soit une ressource adéquate à ce type d'applications. Au contraire, ce besoin de distinctions fines entre concepts n'apparaît pas dans des applications comme l'Indexation Sémantique ou Conceptuelle, la Désambiguïsation Sémantique, la Traduction Automatique et la Recherche d'Information (Mihalcea et Moldovan, *ibid*. ; Krovetz et Croft, 1992 ; Dolan, 1994). Ainsi, Krovetz et Croft (*ibid*.) doutent de l'utilité d'informations de sémantique lexicale de très haute qualité dans le cadre de la Recherche d'Information, car une désambiguïsation implicite s'effectue lorsque de nombreuses correspondances sont établies entre mots d'une requête et mots d'un document. De même, dans le cas de la TA, l'existence d'ambiguïtés parallèles entre les langues réduit généralement la nécessité d'établir des distinctions sémantiques fines au niveau des mots mis en correspondance.

2.3.3.3. Regroupement des sens fournis par des ressources prédéfinies

L'inadéquation de ressources sémantiques très raffinées à certaines applications a généré le développement de méthodes permettant le regroupement de sens lexicaux très proches et la création de ressources de granularité plus grossière issues de ressources existantes (Dolan, 1994 ; Peters *et al*., 1998 ; Mihalcea et Moldovan, 2001 ; Navigli, 2006 ; Navigli *et al*., 2007). Ces méthodes sont d'ailleurs appelées par Dolan **méthodes d'ambiguïsation** (*ambiguation*). Il s'agit en effet de méthodes de **clustering** qui

visent à regrouper les sens similaires des mots et qui ont, pour la plupart, été développées par référence à WordNet[180].

Certaines méthodes exploitent l'organisation hiérarchique de WordNet, c'est-à-dire les relations entre les synsets et les informations lexicales décrites (Peters *et al*., *ibid*. ; Mihalcea et Moldovan, *ibid*.), tandis que d'autres méthodes créent des correspondances entre les sens de WordNet et des définitions sémantiques issues de ressources différentes (Navigli, *ibid*. ; Navigli *et al*., *ibid*.)[181]. Ce type de clustering utilise des critères concernant la similarité des mots inclus dans des synsets décrivant des sens différents du même mot, ou la similarité des relations entretenues entre ces synsets et d'autres synsets du réseau (Mihalcea et Moldovan, *ibid*., Peters *et al*., *ibid*.). Outre l'exploitation des relations entre synsets, Mihalcea et Moldovan font également appel à des critères probabilistes sur la fréquence d'occurrence de synsets dans les textes, fréquence mesurée sur la base d'un corpus sémantiquement étiqueté utilisant WordNet[182]. La réduction de la granularité sémantique opérée par ces méthodes est censée améliorer les résultats des méthodes de désambiguïsation lexicale et de recherche d'information multilingue[183]. Autre avantage découlant de ce clustering sémantique[184], l'amélioration de la compatibilité entre ressources différentes et la facilitation de la mise en correspondance entre des sens décrits.

[180] A l'exception de la méthode de Dolan (*ibid*.), qui concerne les distinctions sémantiques du 'LDOCE'.

[181] Dans cette méthode, les sens de WordNet sont mis en correspondance avec les sens de granularité grossière du dictionnaire 'Oxford English Dictionary' (OED). Le repérage de ces sens dans l' 'OED' est possible grâce à la structure hiérarchique des sens fournis dans WordNet.

[182] Il s'agit du corpus SemCor, extrait du 'Brown corpus', où les mots pleins sont étiquetés à la main par des sens fournis dans WordNet (Miller *et al*., 1993).

[183] Il a été démontré que, dans la Recherche d'Information Multilingue (qui est l'application principale envisagée pour EuroWordNet), l'enrichissement de l'ensemble original des mots polysémiques contenus dans les requêtes avec des synonymes de leurs synsets dans WordNet augmente de manière exponentielle la quantité de bruit dans les résultats des requêtes (Peters *et al*., 1998).

[184] Le clustering des sens se base sur trois types de relations sémantiques. Tout d'abord, la granularité très fine des distinctions sémantiques de WordNet conduit à une prolifération de distinctions sémantiques ayant un haut degré de similarité. Ce recouvrement sémantique rend possible la généralisation à un ensemble de sens, qui constitue un dénominateur commun sous-spécifié que tous les sens partagent. Le clustering peut aussi exploiter la relation de métonymie, qui caractérise les cas de polysémie régulière ou systématique. Une extension métonymique du sens peut souvent être conçue comme une dérivation d'un sens de base. Ainsi, le phénomène de métonymie peut être considéré comme un principe sous-jacent de la structure du lexique, pouvant être exprimé à l'aide de règles lexicales. Un autre type de régularité sémantique qui peut être exploité dans le cas des verbes, est lié au phénomène de l'altération de diathèse, en se fondant sur

L'intérêt des mécanismes de clustering des sens a été démontré au sein du projet **EuroWordNet**[185] (Vossen, 1999), où le problème d'incompatibilité entre ressources était particulièrement évident. Ce projet impliquait la combinaison de WordNets de langues différentes pour la création d'une base de connaissances multilingue. Dans cette base, des sens équivalents entre langues sont liés au sein d'un Index Inter-Langues (*Interlingual Index* en anglais (ILI)). ILI forme le sur-ensemble des concepts rencontrés dans les langues et chaque synset dans les wordnets monolingues possède au moins une relation d'équivalence avec une entrée d'ILI. Des synsets spécifiques aux langues et liés à la même entrée devraient être conceptuellement équivalents. Cependant, cet appariement des wordnets s'avère difficile en raison des divergences existant au niveau de la différentiation des sens. Le degré élevé de différenciation sémantique du WordNet anglais explique que les sens équivalents des wordnets ne sont pas liés au même sens de l'équivalent de traduction anglais mais à des concepts ILI distincts, reflétant des sens différents du même mot. A cause du développement indépendant des wordnets propres aux langues, il a fallu développer des mécanismes, comme le clustering des sens, qui assurent leur compatibilité lors de leur intégration. La création d'entrées inter-langues de granularité grossière, regroupant les concepts ILI, a permis d'éviter les correspondances divergentes entre wordnets locaux et concepts ILI.

2.3.3.4. Absence de liens sémantiques explicites

Outre les problèmes de complexité de traitement et d'incompatibilité, une autre faiblesse des ressources à granularité de description fine concerne l'incapacité d'abstraction des systèmes automatiques, qui ne réussissent à combiner des informations que si leurs liens sont explicitement décrits. Comme nous l'avons déjà mentionné plus haut, un utilisateur humain, au lieu de traiter comme complètement distincts les sens décrits pour un mot dans ce type de dictionnaire, aboutit à une notion plus abstraite de son sens en combinant des

l'hypothèse que les caractéristiques sémantiques des verbes sont reflétées de manière systématique dans les configurations syntaxiques dans lesquelles ils apparaissent (Peters *et al*., 2000).

[185] Dans le cadre d'EuroWordNet, des wordnets ont été développés pour les langues suivantes : le néerlandais, l'italien, l'espagnol, l'anglais, l'allemand, le français, l'estonien et le tchèque.

informations de sens distincts. Cette différence justifie, d'une certaine manière, la difficulté à exploiter des ressources élaborées pour des humains dans un cadre automatique.

L'absence de tels liens est imputable, en partie, à l'incapacité des techniques de représentation des sens utilisées dans la pratique lexicographique à justifier une distinction entre types d'ambiguïté (Dolan, 1994 ; Pustejovsky, 1995 : 29). Les liens pouvant exister entre les sens lexicaux ne sont donc pas décrits au sein des dictionnaires. Le WordNet, par exemple, en tant que lexique d'énumération de sens, ne contient pas de connexions explicites entre les sens. Plus précisément, des reproches adressées à WordNet concernant les liens entre les informations décrites portent sur l'absence de connexions entre les hiérarchies nominales et verbales et entre mots du même topic, l'absence de certaines relations entre sens lexicaux ainsi que de relations morphologiques et actancielles et de restrictions de sélection (Harabagiu *et al*., 1999)[186].

2.3.4. Ressources monolingues dans un cadre bi- (ou multi-)lingue

Un autre point important qu'il faut souligner, concerne la possibilité d'utilisation au sein d'applications multilingues de ressources exploitées pour la désambiguïsation dans un cadre monolingue. La mise en correspondance des sens lexicaux décrits au sein d'une ressource monolingue avec les équivalents de traduction des mots dans une autre langue ne peut se faire de manière directe, en raison de la haute granularité des distinctions sémantiques des ressources monolingues. Autre raison très importante de cette difficulté, la polysémie divergente des mots des langues différentes. Les expériences menées par Mihăltz (2005) et Specia *et al*. (2006b), qui impliquent la mise en relation des sens d'un ensemble de mots décrits dans WordNet avec leurs équivalents de traduction en hongrois et en portugais, ont montré que la plupart des sens des mots anglais étudiés ont des traductions communes avec d'autres sens, c'est-à-dire qu'une traduction couvre plus de deux sens. La désambiguïsation entre un grand nombre de sens n'est donc pas nécessaire pour la traduction et peut même

[186] Harabagiu *et al*. (*ibid*.) soulignent, en outre, l'absence de certains concepts (sens lexicaux), la non-uniformité et l'inconsistance des définitions dues à leur création manuelle.

provoquer des erreurs si le faux sens est sélectionné – ce qui démontre la granularité trop élevée de WordNet pour pouvoir être appliqué efficacement à la traduction entre l'anglais et les deux autres langues en question.

Un autre problème lié à l'utilisation de méthodes monolingues pour la désambiguïsation dans un cadre multilingue, mis en évidence par ces expériences, concerne la correspondance entre un sens de la LS et plusieurs équivalents de traduction. Dans ce cas, la non désambiguïsation du mot source pendant la traduction, dans la mesure où celui-ci n'est pas ambigu dans la LS, peut provoquer de graves erreurs.

L'exploitation de ce type de ressources dans un cadre multilingue peut donc générer un travail inutile, voire même des erreurs de désambiguïsation. L'adoption de stratégies spécifiques pour obtenir de bons résultats dans un tel cadre a été proposée. Une de ces stratégies consiste à utiliser des correspondances entre plusieurs sens et un équivalent pour réduire l'ambiguïté des mots de la LS et créer un inventaire de sens de granularité plus grossière, où un plus petit nombre de sens davantage distincts devraient être discriminés (Miháltz, 2005.) Cette suggestion pourrait être caractérisée comme une technique d' « ambiguïsation » des sens de WordNet, à l'instar de celles proposées par Dolan (1994), Peters *et al*. (1998) et Mihalcea et Moldovan (2001) (cf. §2.3.3.3).

Les inconvénients liés à l'utilisation de ressources préétablies pour la désambiguïsation, que nous venons d'exposer, expliquent le développement de méthodes ne nécessitant pas d'inventaires de sens prédéfinis. Ces méthodes utilisent des inventaires construits à partir de données et sont ainsi caractérisées comme des méthodes dirigées par les données. La méthode de désambiguïsation que nous proposons dans ce travail est également une méthode dirigée par les données. Nous allons désormais décrire les aspects caractéristiques les plus importants des méthodes existantes et présenter les principes régissant leur fonctionnement.

3. Désambiguïsation lexicale dirigée par les données

3.1. Apprentissage automatique pour la désambiguïsation

3.1.1. Exploitation d'informations extraites de corpus textuels

Dans les méthodes de désambiguïsation **dirigées par les données**, les informations nécessaires à la désambiguïsation proviennent de textes réels, ce qui élimine le besoin de recourir à des ressources préétablies. Ces informations peuvent être repérées manuellement ou de manière automatique. Dans la méthode proposée par Kelly et Stone (1975), par exemple, les règles de désambiguïsation d'un grand nombre de mots sont élaborées manuellement à partir de concordances. Les indices pour la désambiguïsation sont extraits du contexte local des mots et concernent leurs collocations, leurs relations syntaxiques et leur appartenance à des catégories sémantiques communes. Néanmoins, l'élaboration manuelle de règles de désambiguïsation est une entreprise qui requiert beaucoup de temps et est difficilement paramétrable pour de nouvelles langues. Cette difficulté à construire manuellement un grand nombre de règles de désambiguïsation a suscité le développement de méthodes automatiques.

Les méthodes automatiques de désambiguïsation lexicale dirigées par les données se basent sur des techniques d'**apprentissage automatique**, **supervisé** ou **non supervisé**.

3.1.2. Méthodes supervisées de désambiguïsation lexicale

L'apprentissage supervisé présuppose l'existence d'une base de **données d'apprentissage** contenant des exemples de cas déjà traités. Dans ces cas, les résultats possibles sont connus à l'avance et les algorithmes doivent apprendre à combiner une entrée particulière à un résultat. Les données d'apprentissage consistent en des paires d'objets d'entrée et de sortie et l'apprentissage permet la création d'une fonction à partir de ces données. La tâche de l'apprenant

supervisé est de prédire la valeur de la fonction pour chaque objet d'entrée valable, après avoir analysé l'ensemble des exemples d'apprentissage et avoir établi des généralisations à partir des données pour des cas non rencontrés. La sortie de la fonction concerne la prédiction d'une étiquette de classe pour l'objet d'entrée, ce qui rapproche l'apprentissage supervisé d'une **tâche de classification**[187].

Dans le cas de la désambiguïsation lexicale, l'entrée est constituée par une nouvelle instance d'un mot ambigu et ses traits, tandis que la sortie est le sens correct véhiculé par cette instance du mot.

Les techniques d'apprentissage supervisé s'appuient sur un ensemble d'apprentissage réunissant des exemples d'instances de mots ambigus qui ont déjà été désambiguïsées[188]. Ces données sont fournies sous la forme de **corpus sémantiquement étiquetés** (Weiss, 1973 ; Black, 1988 ; Leacock *et al*., 1993). Lors de l'étape d'apprentissage, les techniques supervisées (arbres de décision, réseaux de neurones, méthodes basées sur les probabilités, etc.) apprennent à associer des ensembles de traits des mots à un sens particulier issu d'une liste de sens fournis. L'objectif est de prédire, après analyse des exemples, le sens correct pour de nouvelles instances des mots ambigus et, éventuellement, de leur attribuer une étiquette sémantique.

Les étiquettes sémantiques employées correspondent généralement aux sens fournis par une ressource préétablie, par exemple les sens décrits dans WordNet (Leacock *et al*. 1993 ; Ng et Lee, 1996) ou ceux décrits dans le LDOCE (Bruce et Wiebe, 1994a,b ; Pedersen *et al*., 1997 ; Pedersen et Bruce, 1997b). L'étiquetage manuel des corpus nécessite beaucoup de temps, ce qui explique la faible quantité de corpus sémantiquement étiquetés. Cette difficulté à étiqueter manuellement constitue donc un obstacle à l'acquisition de connaissances lexicales à partir de corpus.

Des tentatives d'étiquetage automatique utilisant des **méthodes d'amorçage** (*bootstrapping*) ont alors vu le jour (Hearst, 1991 ; Yarowsky, 1995 ; Basili *et al*.,

[187] Dans une tâche de classification, des éléments individuels sont regroupés sur la base d'informations concernant une ou plusieurs caractéristiques inhérentes à ces éléments et d'un ensemble d'apprentissage constitué d'éléments étiquetés à l'avance.

[188] Certaines techniques utilisées sont les arbres de décision (Black, 1988), les classificateurs de Bayes (Gale *et al*., 1992 ; Leacock *et al*. 1993) et les réseaux de neurones.

1997). Ces méthodes impliquent une phase d'apprentissage (ou d'entraînement) sur un petit ensemble d'instances de mots désambiguïsées et étiquetées manuellement du point de vue sémantique. Les informations statistiques extraites du contexte des instances des mots pendant l'étape d'apprentissage sont ensuite utilisées pour en désambiguïser d'autres. Lorsqu'une nouvelle instance est désambiguïsée avec certitude, le système acquiert automatiquement des informations statistiques additionnelles et améliore, de cette manière, ses connaissances de façon incrémentale.

D'autres solutions ont été proposées au problème de l'étiquetage manuel des données d'entraînement et concernent, cette fois-ci, l'utilisation de **corpus bilingues parallèles** (Brown *et al*., 1991b ; Gale *et al*., 1992a, 1993 ; Resnik, 2004), où les mots d'une langue sont étiquetés par leurs équivalents de traduction mis en évidence par un processus d'alignement lexical ou par l'utilisation de corpus monolingues combinés à des dictionnaires bilingues (Dagan *et al*., 1991, 1994).

3.1.3. Méthodes non supervisées de désambiguïsation lexicale

Pour les méthodes non supervisées de désambiguïsation lexicale, dirigées par les données (Schütze, 1998 ; Pedersen et Bruce, 1997a, 1998 ; Véronis, 2003, 2004 ; Bruce et Wiebe, 1994a, 1994b), les données préétiquetées sont inutiles. Les connaissances nécessaires à la désambiguïsation sont automatiquement identifiées dans les textes traités. Les sens possibles des mots ambigus sont repérés dans les textes en regroupant les instances des mots sur la base de traits contextuels divers. Des processus d'acquisition automatique de sens ont déjà été présentés dans le paragraphe 1 du chapitre 2.

L'analyse du contexte, effectuée pour la détermination des sens des mots ambigus, permet le repérage des traits avec lesquels le contexte de nouvelles instances des mots sera comparé, par la suite, pour la désambiguïsation. Dans la méthode de Schütze (*ibid*.), par exemple, où les sens lexicaux correspondent à des clusters de vecteurs contextuels, la désambiguïsation d'une nouvelle instance d'un mot s'opère en comparant le vecteur construit pour le nouveau contexte avec la centroïde de chaque cluster (la moyenne de ses éléments) et en sélectionnant ensuite le cluster dont la centroïde est la plus proche du vecteur

contextuel. Le cluster retenu correspond au sens du mot dans le nouveau contexte.

En revanche, dans le travail de Pedersen et Bruce (*ibid*.), l'étape d'acquisition de sens coïncide avec celle de désambiguïsation. Le processus de désambiguïsation est appliqué sur un corpus sémantiquement étiqueté à des fins d'évaluation. Les étiquettes sémantiques ne sont pas utilisées lors de l'apprentissage (qui est non supervisé) mais servent à l'évaluation des groupes de sens générés, mis en correspondance avec les étiquettes.

3.2. Impact de la dispersion des données

Outre le manque de ressources sémantiquement annotées, il existe un autre obstacle à l'acquisition à partir de corpus des connaissances lexicales nécessaires au fonctionnement des méthodes de désambiguïsation lexicale dirigées par les données, supervisées ou non. Cet obstacle consiste en la **dispersion des données**. La quantité de textes nécessaire pour assurer la représentation de la totalité des sens des mots polysémiques est énorme, étant donné les différences importantes qui existent entre la fréquence des sens lexicaux. Les cooccurrences possibles d'un mot polysémique sont par ailleurs très nombreuses et difficiles à rencontrer, même dans un corpus très large, où il se peut qu'elles apparaissent trop peu fréquemment pour être significatives (Ide et Véronis, 1998).

Une solution au problème de la dispersion des données consiste à utiliser des **modèles basés sur les classes**, qui essaient d'obtenir de meilleures estimations en combinant les observations de classes de mots conçus comme appartenant à la même catégorie. Brown *et al*. (1990b), Pereira et Tishby (1992) et Pereira *et al*. (1993) proposent des méthodes qui dérivent des **classes distributionnelles** du corpus ; Lyse (2006) enrichit le corpus par des **classes de sens** constituées à partir des résultats des Miroirs Sémantiques ; d'autres auteurs utilisent, quant à eux, des sources d'informations externes pour la définition des classes de mots. Resnik (1992) exploite WordNet, Yarowsky (1992) utilise les catégories de Roget's, tandis que Slator (1992) se sert des codes de domaine de 'LDOCE'. Les méthodes basées sur les classes répondent en partie au problème de la dispersion des données et éliminent le besoin de recourir à des données

étiquetées à l'avance. Leur inconvénient principal est néanmoins de provoquer une **perte d'informations** en raison de l'hypothèse, très forte, selon laquelle tous les mots inclus dans la même classe se comportent de façon similaire.

D'autres méthodes se basent sur la similarité des **motifs de cooccurrence**, sans pour autant former des classes de mots (Dagan *et al*., 1993 ; Dagan *et al*., 1994 ; Grishman et Sterling, 1993). L'estimation de la probabilité d'une cooccurrence de mots jamais rencontrée repose sur des données à propos de cooccurrences observées dans le corpus contenant des mots similaires. La performance de ce type de méthodes est considérée comme meilleure que celle des méthodes basées sur les classes.

Il existe aussi des méthodes qui combinent les deux approches. Par exemple, la méthode de Black (1998) utilise trois types différents de catégories contextuelles : les catégories de domaine du 'LDOCE', un ensemble de cooccurrents apparaissant très fréquemment à des positions très proches et un ensemble de cooccurrents apparaissant à des positions plus éloignées du mot ambigu, au sein de la ligne de concordance.

4. Désambiguïsation lexicale orientée vers des applications précises

4.1. Désambiguïsation lexicale : une étape intermédiaire de traitement

La plupart des méthodes de désambiguïsation présentées jusqu'ici sont indépendantes d'applications particulières. Ce phénomène est pourtant contradictoire avec la nature de la tâche de désambiguïsation, constituant une étape intermédiaire de traitement dont le but est d'améliorer la performance de certaines applications du TAL (Wilks et Stevenson, 1996). Les besoins de désambiguïsation dans le cadre d'applications différentes divergent, ainsi que la possibilité d'exploiter ses résultats (Resnik et Yarowsky, 1997 ; Mihalcea et Moldovan, 2001 ; Krovetz et Croft, 1992 ; Dolan, 1994). La désambiguïsation est donc une tâche fortement dépendante des applications et dont la réussite dépend de leurs besoins particuliers. Cette absence de lien entre étape de

désambiguïsation et finalité de l'application influence considérablement l'applicabilité des méthodes proposées dans des cadres différents et a provoqué l'apparition de méthodes orientées vers des applications particulières.

4.2. Désambiguïsation lexicale pour la traduction

Nous allons ici nous intéresser essentiellement aux méthodes de désambiguïsation opératoires dans le cadre de la TA (Kaji et Morimoto, 2002 ; Kaji *et al*., 2003 ; Vickrey *et al*., 2005 ; Specia, 2005 ; Mihăltz, 2005 ; Specia *et al*., 2006a,c, 2007). Certaines de ces méthodes sont des **méthodes hybrides**, qui combinent des informations trouvées dans des ressources prédéfinies à des informations repérées dans les corpus (Specia, 2005 ; Mihăltz, 2005). D'autres sont basées seulement sur des informations extraites à partir de textes qui consistent, le plus souvent, en des informations contextuelles provenant des deux parties de corpus parallèles alignés (bitextes) et des informations de traduction.

De manière générale, dans les méthodes de désambiguïsation développées pour être utilisées dans un cadre de traduction, la tâche de désambiguïsation est étroitement liée à celle de sélection lexicale. L'**assimilation** des deux tâches est parfois telle que, dans certains travaux, les prédictions concernant les sens véhiculés par les mots ambigus sont directement considérées comme des prédictions de leur traduction dans une autre langue (Vickrey *et al*., 2005). Concevoir la tâche de **désambiguïsation lexicale** comme une tâche de **traduction lexicale** présente certains avantages, comme la grande disponibilité de données « étiquetées »[189] et la possibilité d'éviter le repérage de distinctions entre nuances sémantiques fines, étape qui, nous le rappelons, n'est pas toujours nécessaire, étant donné les ambiguïtés parallèles entre langues.

Les méthodes de désambiguïsation développées dans un cadre de traduction combinent généralement des informations contextuelles à des informations traductionnelles. Les informations contextuelles proviennent d'un ou des deux côtés de corpus parallèles. Dans la méthode de Specia *et al*. (2006a, 2007), la désambiguïsation de nouvelles instances de mots polysémiques source

[189] Dans le cas de corpus bilingues alignés au niveau des mots.

s'opère en utilisant des informations de cooccurrence provenant du contexte de la traduction dans la LC, c'est-à-dire du texte déjà traduit. Ces informations sont utilisées pour créer des *n*-grammes et des requêtes sous forme de « sac de mots » pour chaque traduction possible du mot source, incluant la traduction en question et ses cooccurrents dans le texte déjà traduit. Ces requêtes servent ensuite à analyser la fréquence d'apparition des traductions possibles des mots polysémiques dans des fragments de textes de la LC trouvés sur le Web. La traduction sélectionnée est celle qui apparaît dans les requêtes pour lesquelles le nombre maximal de résultats est obtenu (selon 'Google'). Néanmoins, la remarque des auteurs selon laquelle, malgré les bons résultats de la méthode, il serait préférable de la considérer comme une source complémentaire d'informations pour la désambiguïsation lexicale – pouvant être combinées avec des informations provenant d'autres sources[190] – nous paraît tout à fait justifiée. Dans la méthode proposée par Kaji et Morimoto (2002) la désambiguïsation s'effectue sur la base de données de corrélation de type 'sens vs indice', acquises pendant l'étape de repérage de sens dans la LS. Lors d'une nouvelle instance du mot source, un score est calculé pour chacun de ses sens possibles, qui correspond à la somme des corrélations entre le sens et les indices trouvés dans le contexte. Le sens sélectionné est alors celui qui obtient le score le plus élevé.

Après avoir exposé un ensemble de méthodes de désambiguïsation lexicale basées sur les connaissances et dirigées par les données, dépendantes et indépendantes d'applications précises, nous allons, dans le paragraphe suivant, analyser certains aspects liés à l'évaluation de leurs résultats.

[190] Les autres sources d'informations pourraient concerner le contexte de la LS (parties du discours, relations syntaxiques et collocations) (Specia, 2005).

5. Evaluation des méthodes de désambiguïsation lexicale

5.1. Nécessité d'un standard commun pour l'évaluation

Comme nous l'avons déjà mentionné plus haut, la désambiguïsation lexicale constitue une **tâche intermédiaire** dans le TAL et non un but en soi. « Intermédiaire » signifie ici qu'il s'agit d'une tâche dont l'évaluation est déterminée par des **critères linguistiques** ou **théoriques**, contrairement à des tâches comme la traduction automatique ou la recherche d'information, dont la qualité peut être estimée par les utilisateurs finaux (Wilks et Stevenson, 1996).

Cette distinction, difficile à établir clairement, est pourtant essentielle. Elle correspond à l'intuition que l'on n'a pas besoin d'étiquetage morphosyntaxique, d'analyse syntaxique ou de désambiguïsation lexicale en tant que tels, mais seulement comme d'un moyen pour parvenir à un certain but. Le seul cas où les informations intermédiaires de ces types deviennent essentielles en soi est celui où le but consiste à vérifier ou à réfuter une théorie de traitement ou de structure linguistique[191].

L'évaluation des méthodes de désambiguïsation nécessite donc auparavant la définition de critères permettant l'estimation de leur performance et l'appréciation de la qualité des résultats qu'elles fournissent. Des tentatives de définition de critères et de standards, permettant la comparaison des résultats de différentes méthodes de désambiguïsation, ont vu le jour assez « tardivement » par rapport à d'autres tâches du TAL, comme l'analyse syntaxique et l'étiquetage morphosyntaxique[192]. Cette absence de standard commun pour l'évaluation, conjuguée aux divergences des approches de désambiguïsation proposées (tant au niveau des ensembles de mots polysémiques étudiés qu'à celui des corpus et

[191] D'après Wilks et Stevenson (*ibid.*), cette différence quant à l'objectif des deux types de tâche, couplée à la préoccupation des chercheurs de confirmer ou de réfuter des théories – préoccupation plus importante que le souci de fournir des résultats utilisables –, a permis aux tâches intermédiaires de devenir très importantes dans le domaine du TAL, parfois même aux dépenses des tâches « finales ».

[192] Des ressources standardisées communes, disponibles pour l'apprentissage et l'évaluation sont, par exemple, le Penn Treebank (Marcus *et al.*, 1993) pour l'analyse syntaxique, et les versions annotées de Brown (Kucera et Francis, 1967) et de Lancaster-Oslo-Bergen corpus (Johansson, 1980 ; Johansson et Hofland, 1987) pour l'étiquetage morphosyntaxique.

des sources d'informations utilisées[193]), ont provoqué une multiplication des méthodes d'évaluation, dont le nombre équivaut à celui des méthodes de désambiguïsation. L'uniformisation de l'évaluation de ces méthodes nécessiterait, d'une part, un ensemble de **métriques d'évaluation** permettant la comparaison des résultats de méthodes différentes et, d'autre part, la disponibilité de **données sémantiquement étiquetées** de bonne qualité. Ces données pourraient servir à l'entraînement et à l'évaluation d'algorithmes d'apprentissage, utilisés par les méthodes supervisées de désambiguïsation. Elles pourraient, en outre, constituer des ensembles de test permettant l'estimation rigoureuse de la performance des algorithmes ainsi que la comparaison de leur performance. Il faut néanmoins préciser que de telles ressources sont difficiles à obtenir.

5.2. Campagnes d'évaluation des systèmes de désambiguïsation lexicale

Dans le domaine de la désambiguïsation lexicale, des exercices d'évaluation ont été entrepris, comme le SENSEVAL (Kilgarriff, 1998a ; Kilgarriff et Rosenzweig, 2000a,b ; Kilgarriff, 2002) et le ROMANSEVAL (Calzolari et Corazzari, 2000 ; Segond, 2000). Dans le cadre de ces exercices, les systèmes supervisés s'entraînent sur un même ensemble de données d'apprentissage (partie d'un corpus qui sert d'étalon d'or). Un ensemble commun de données (l'autre partie de l'étalon d'or) est également utilisé pour l'évaluation des systèmes participant à une tâche précise (supervisés ou non), ce qui rend possible la comparaison des résultats de la désambiguïsation.

Resnik et Yarowsky (1997 ; 2000) expriment un ensemble de suggestions très pertinentes à propos des métriques d'évaluation des algorithmes de désambiguïsation, de la construction d'un ensemble de test et de l'adoption d'un jeu d'étiquettes sémantiques. Leur proposition, quant aux **métriques d'évaluation**, concerne la prise en compte de la distance et des relations entre les sens, paramètres qui justifieraient une pénalisation des erreurs de

[193] Sources qui vont d'indices collocationnels locaux (Yarowsky, 1993) à l'appartenance des mots à des classes de mots sémantiquement liés ou liés par le topic (Gale *et al*., 1992).

désambiguïsation adaptée à ces relations inter-sens[194]. Cette **pénalisation variable** est implicitement liée aux possibilités de mésinterprétation, dans la mesure où le risque augmente proportionnellement à la distance entre le sens véhiculé par une nouvelle instance du mot et le sens sélectionné par l'algorithme de désambiguïsation.

En ce qui concerne la création d'un **ensemble de test** commun, tout en prenant en considération les besoins différents des méthodes supervisées et non-supervisées en matière de quantité et de qualité des données, les auteurs proposent un cadre qui combine une large couverture à la possibilité d'évaluation par référence à un ensemble restreint de mots.

Quant au sujet de l'ensemble d'**étiquettes sémantiques** qui pourrait être utilisé dans un cadre commun d'évaluation, Resnik et Yarowsky suggèrent de restreindre l'inventaire des sens des mots d'une langue aux distinctions sémantiques lexicalisées dans d'autres langues[195]. Cette solution se situe à mi-chemin entre les distinctions au niveau des homographes et l'expression de toutes les distinctions de granularité très fine des dictionnaires monolingues. Les distinctions inter-langues pourraient, selon les auteurs, être également liées aux numéros de sens au sein de ressources lexicales (comme WordNet et 'LDOCE'), afin de créer une source de référence pour l'apprentissage et l'évaluation.

[194] Le choix du sens erroné d'un homographe (par ex. le sens « rive » du mot *bank* pour une instance véhiculant le sens « institution financière ») devrait être davantage pénalisé que la sélection d'un sens proche de celui effectivement véhiculé par la nouvelle occurrence du mot (par ex. le sens « bâtiment »). Les informations sur la distance entre les sens pourraient être repérées sur la base de la distance et de la hiérarchie des sens et des sous-sens dans un dictionnaire ou dans une hiérarchie sémantique (comme WordNet).

[195] L'idée est de définir un ensemble de LC et de conserver les distinctions sémantiques des mots d'une langue lexicalisées dans un sous-ensemble minimal des LC. Cette méthode est adoptée dans la tâche « Multilingual lexical sample » de Senseval-3, qui vise à créer un cadre pour l'évaluation des systèmes de TA en mettant l'accent sur la traduction des mots ambigus. Cette tâche ressemble à la tâche « lexical sample » de Senseval, mais l'inventaire de sens, au lieu d'être extrait d'un dictionnaire, est constitué des traductions des mots ambigus dans une autre langue, d'après la suggestion de Resnik et Yarowsky. Les contextes utilisés sont anglais et les étiquettes pour les mots ambigus sont leurs traductions dans une autre langue. Des mots avec des degrés divers d'ambiguïté inter-langue sont sélectionnés afin d'illustrer de manière complète les problèmes qui peuvent apparaître. Deux paires de langues sont utilisées dans le Senseval-3 : anglais-français et anglais-hindi et 50 mots ambigus sont traités pour chaque paire de langues. Une tâche similaire apparaît dans la campagne d'évaluation SemEval-1/Senseval-4. La tâche « Multilingual Chinese-English Lexical Sample Task » a pour but l'évaluation de systèmes de TA anglais-chinois. Là aussi, les étiquettes sémantiques des mots ambigus chinois correspondent à leurs traductions en anglais.

Les propositions de Resnik et Yarowsky (*ibid*.) ont été reprises pour la création d'un standard facilitant la communication, la collaboration et l'évaluation rigoureuse des méthodes de désambiguïsation, dans le cadre de l'exercice d'évaluation SENSEVAL. Les buts de cette campagne sont justement de mettre en compétition différents systèmes de désambiguïsation, de les comparer, d'évaluer leur performance et d'estimer leurs forces et leurs faiblesses relativement à des mots, des variétés linguistiques et des langues différentes.

5.3. Evaluation de la désambiguïsation par rapport au résultat de la tâche finale

Même si la désambiguïsation lexicale constitue une tâche intermédiaire et non un but en soi, une méthode alternative d'évaluation de la performance des algorithmes consisterait à se référer au résultat de la tâche finale, pour laquelle le module de désambiguïsation est utilisé. Ce résultat pourrait être, soit évalué par des humains, soit comparé aux résultats fournis par le même système sans le module de désambiguïsation – ce qui correspondrait à une espèce de **méthode de base** (baseline). Schütze et Pedersen (1995) et Schütze (1998) évaluent, par exemple, la performance d'un algorithme de désambiguïsation par référence aux résultats d'un système baseline de recherche d'informations. L'expérience compare les résultats de la recherche d'informations basée sur les sens à ceux obtenus par un système de recherche d'informations basée sur les mots. Les documents et les requêtes sont représentés en tant que vecteurs dans un espace multidimensionnel dont chaque dimension correspond à un mot (lors d'une recherche basée sur les mots), et à un sens (lors d'une recherche basée sur les sens)[196].

Dans le cas de la Traduction Automatique, les traductions fournies par un système statistique qui n'utilise pas de module de désambiguïsation lexicale (système « baseline ») peuvent être comparées aux traductions fournies par le système à partir des mêmes textes source, lorsque ce type de module y est

[196] Les mots sont désambiguïsés en utilisant la discrimination de groupes de contextes. Les documents et les requêtes dans lesquels un mot attribué à un sens particulier apparaît, ont une valeur différente de zéro au niveau de la dimension correspondante.

intégré. Cette méthode est employée dans les travaux de Carpuat et Wu (2005a), de Cabezas et Resnik (2005) et de Chan *et al*. (2007). Il faut pourtant souligner que cette manière d'évaluer les algorithmes de désambiguïsation se heurte aux inconvénients des métriques d'évaluation utilisées pour les différentes tâches. Par exemple, la métrique BLEU (Papineni *et al*., 2002), utilisée pour l'évaluation des systèmes de traduction, est souvent critiquée comme ne prenant en compte que les correspondances exactes entre les traductions proposées et les traductions de référence (plus précisément, des correspondances de *n*-grammes) ; ce qui empêche la prise en compte d'une proposition sémantiquement correcte ne correspondant pas exactement à la traduction de référence (Callison-Burch *et al*., 2006a,b). Nous reviendrons sur les inconvénients de cette métrique d'évaluation dans le chapitre abordant l'intégration de modules de désambiguïsation dans les systèmes de TA (chapitre 8), ainsi que dans celui de d'évaluation de notre propre méthode de désambiguïsation (chapitre 10).

CONCLUSION

Dans ce chapitre, nous avons présenté un ensemble de méthodes de désambiguïsation lexicale basées sur des sources de connaissance externes et sur des résultats de méthodes d'acquisition des sens. Quelle que soit la manière de créer des inventaires sémantiques, la source des informations exploitées pour la désambiguïsation de nouvelles instances des mots est le plus souvent les nouveaux contextes. C'est pourquoi nous avons analysé les différentes conceptions de la notion de **contexte** dans un cadre monolingue et bilingue. A la fin du chapitre, nous avons étudié certains aspects liés à l'évaluation de la performance des méthodes de désambiguïsation et les problèmes qui surgissent lors d'une telle tâche.

Etant donné les inconvénients liés à l'utilisation de sources externes pour la désambiguïsation, à propos tant de leur conformité au traitement automatique que de leur disponibilité pour des domaines et des langues différents, nous allons désormais nous concentrer sur la possibilité d'analyse et de résolution de l'ambiguïté lexicale sur la base d'**informations internes**, c'est-à-dire

d'informations extraites de corpus. Le repérage de sens et la désambiguïsation, dans le cadre de notre travail, seront menés dans un contexte bilingue et auront pour but la création de correspondances d'ordre sémantique entre les mots de deux langues en relation de traduction.

Pour implémenter et évaluer les méthodes d'acquisition de sens, de désambiguïsation et de sélection lexicale proposées, nous avons utilisé deux corpus parallèles différents : un corpus d'apprentissage et un corpus de test. Ces corpus ont dû subir certaines étapes de prétraitement, afin que les informations contenues soient exploitables. Dans le chapitre suivant, nous présenterons en détail ces étapes de prétraitement des corpus d'apprentissage et d'évaluation.

CHAPITRE 4

PRETRAITEMENT DES DONNEES

INTRODUCTION

Les méthodes développées dans le cadre de cette thèse sont des méthodes empiriques, qui utilisent des informations extraites de corpus textuels. Elles sont par ailleurs endogènes, dans le sens où elles ne nécessitent pas l'utilisation de ressources lexicales ou sémantiques prédéfinies et où la totalité des informations nécessaires pour le traitement est repérée dans les corpus. Etant donné que l'acquisition de sens, la désambiguïsation et la prédiction de traduction sont orientées, dans ce travail, vers le traitement dans un cadre bilingue et, plus précisément, dans un but de traduction, les corpus utilisés sont des corpus bilingues parallèles. Ces corpus contiennent des textes originaux et leurs traductions dans une autre langue. Les méthodes proposées ne nécessitant pas d'informations relatives aux langues traitées, elles pourraient théoriquement s'appliquer à n'importe quelle paire de langues. Les expériences menées dans le

cadre de ce travail portent sur la paire de langues anglais-grec ; les corpus bilingues utilisés concernent donc cette paire de langues.

Deux corpus parallèles ont été utilisés : l'un pour l'apprentissage et l'autre pour l'évaluation. Le corpus d'apprentissage sert au repérage des informations qui permettent la création de correspondances sémantiques inter-langues et à la modélisation des paramètres nécessaires pour le traitement (à savoir la désambiguïsation et la prédiction de traduction) de nouveaux cas. Le corpus d'évaluation (ou de test) sert précisément à fournir de nouveaux cas, dont le traitement permet d'évaluer la performance des méthodes proposées.

1. Corpus d'apprentissage

1.1. Caractéristiques du corpus d'apprentissage

Les méthodes d'acquisition de sens, de désambiguïsation et de prédiction de traduction qui ont été implémentées sont des méthodes dirigées par les données. Ces méthodes ne présupposent pas l'utilisation de ressources de connaissances externes. Les informations nécessaires sont recueillies à partir de corpus textuels, lors d'une étape d'apprentissage non supervisé. Le corpus utilisé pour cette tâche est donc appelé **corpus d'apprentissage**. Il ne constitue bien souvent qu'une partie d'un corpus, l'autre partie étant réservée à l'évaluation. Ceci n'est pas le cas dans notre travail, dans la mesure où le corpus utilisé pour l'évaluation (la partie anglais-grec du corpus EUROPARL) diffère du corpus d'apprentissage. Dans ce paragraphe nous allons décrire la nature et les caractéristiques de notre corpus d'apprentissage et nous présenterons ensuite les étapes de son prétraitement.

Le corpus utilisé pour l'apprentissage correspond à la partie anglais-grec du corpus parallèle multilingue **INTERA** (Gavrilidou *et al*., 2004). La partie anglais-grec du corpus comprend environ 4 000 000 mots ; les textes relèvent de cinq domaines différents : droit (42% des textes du corpus), santé (24%), éducation (21%), tourisme (11%) et environnement (2%). Outre le *Journal de l'Union Européenne*, qui constitue la principale source de textes pour les quatre

premiers domaines, les autres sources utilisées pour la constitution du corpus sont les suivantes : le Réseau Judiciaire Européen, la Cour de justice des Communautés européennes et le Ministère grec des Affaires étrangères (droit) ; l'Office national hellénique du Tourisme (tourisme) ; l'Institut Pasteur d'Athènes (santé) ; le Centre national d'orientation professionnelle (EKEP), l'Agence européenne pour le développement de l'éducation des personnes ayant des besoins particuliers et le Réseau d'information sur l'éducation en Europe (éducation) ; l'Organisation Internationale 'Biopolitics' (environnement).

Ce corpus parallèle présente la particularité que la direction de traduction n'est pas la même pour tous les textes. Il s'agit d'un corpus bi-directionnel (Altenberg et Granger, 2002b), qui contient des textes originaux dans les deux langues (anglais et grec) et leurs traductions dans l'autre. Cette particularité du corpus pourrait avoir un impact significatif sur les résultats d'une analyse. L'existence de caractéristiques propres aux textes traduits ayant été démontrée à plusieurs reprises (Baker, 1993, 1995 ; Mauranen, 2002 ; Frawley, 1984b), il serait par conséquent important que la nature des textes soit prise en compte lors de l'étude des phénomènes révélés par ce corpus[197]. La possibilité de procéder à une analyse par type de textes (originaux et traductions) rendrait alors possible la comparaison des résultats obtenus et l'étude de leurs divergences, et permettrait ainsi d'aboutir à des conclusions probablement intéressantes à propos des caractéristiques des textes traduits, qui les différencient des textes originaux[198].

Le problème qui se pose dans le cas de notre corpus et qui empêche toute séparation entre textes originaux et traductions dans les deux langues ainsi que la comparaison des résultats de l'analyse sur les deux types de texte, est que la direction grec-anglais est assez « défavorisée ». Ainsi, des 534 textes originaux, seuls 106 d'entre eux (soit 19,85 %) sont des originaux grecs, les 428 autres textes étant des originaux anglais (80,14 %). Une tentative de division du corpus en deux selon la nature des textes a bien été entreprise, mais il a été tout de suite

[197] La nécessité de prendre en compte ce paramètre n'est pas valable pour toutes les études traductionnelles. Par exemple, dans la méthode d'analyse sémantique 'Miroirs Sémantiques' (Dyvik, 2003, 2005), la traduction est considérée comme symétrique. Cette méthode sera présentée en détail dans le chapitre 6.

[198] L'analyse de ces résultats permettrait l'élaboration d'hypothèses sur les facteurs qui interviennent lors des choix lexicaux dans les traductions, et qui diffèrent probablement de ceux qui déterminent l'usage des mots dans des textes originaux.

évident que la différence d'étendue des deux parties aurait eu une influence conséquente sur l'analyse, rendant même impossible la comparaison entre les résultats obtenus par chacune des deux parties. Plus concrètement, l'un des objectifs de l'apprentissage dans le cadre de ce travail est l'analyse sémantique d'un ensemble de mots anglais au moyen d'informations relatives aux mots grecs qui leur correspondent du point de vue de la traduction (partie gauche de la figure 1). Ces informations relatives aux mots grecs proviennent des contextes des instances du mot anglais source qui leur correspondent[199]. L'utilisation pour la même tâche des textes originaux grecs et de leurs traductions impliquerait l'exploitation d'informations liées aux mots grecs source, issus des textes de la LS et traduits par le mot polysémique anglais en question, pour l'analyse de la sémantique du mot anglais (cf. partie droite de la figure 1). Dans ce cas, les informations relatives aux mots grecs proviendraient des contextes correspondants aux instances de leur équivalent de traduction, dans les textes de la LC.

Si la comparaison des résultats de ces deux démarches était possible, nous pourrions éventuellement aboutir à des conclusions concernant l'apport respectif des contextes des mots dans des textes originaux (premier cas) et dans des textes traduits (deuxième cas) pour l'analyse de leur sémantique. Néanmoins, le petit volume de textes originaux grecs ne permet pas le repérage d'instances pour l'ensemble des mots grecs qui sont des équivalents possibles du mot anglais source[200]. De plus, les instances du mot anglais dans les traductions ne véhiculent qu'une faible partie des sens qu'il véhicule en tant que mot source dans le reste du corpus.

[199] Nous expliquerons en détail cette procédure dans le paragraphe 1.3. du chapitre 6.

[200] Dans la figure, tous les équivalents grecs sont retrouvés comme mots source dans les textes originaux grecs traduits par le mot anglais dans la LC. Cette représentation sert à illustrer l'utilisation des équivalents en tant que mots source mais, dans la pratique, cette symétrie parfaite ne se retrouve pas.

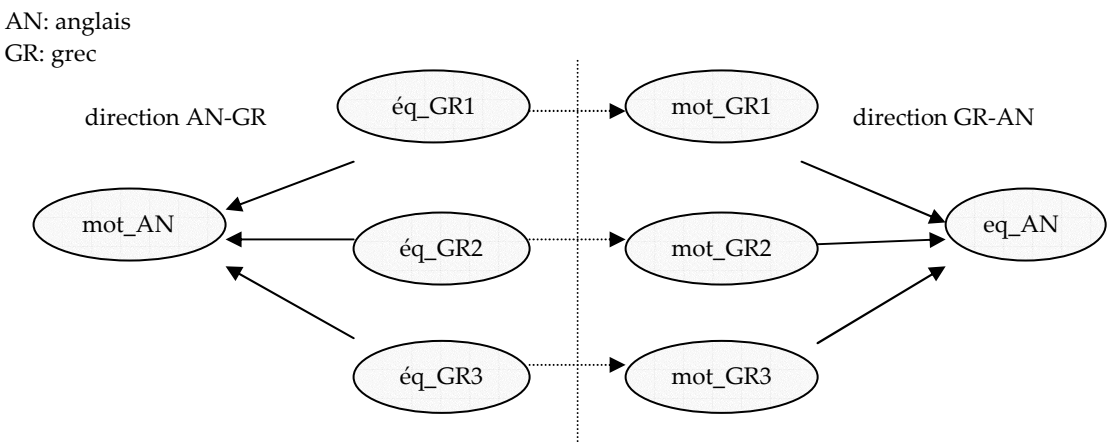


Figure 1. Correspondances utilisées au sein des sous-corpus de textes originaux anglais et grecs

Face à ce faible nombre de textes originaux grecs, nous avons tenté de mener notre analyse sur deux étapes : la première étape impliquait les informations contenues dans l'ensemble du corpus et la deuxième impliquait seulement celles qui se trouvent dans la partie du corpus constituée des textes originaux anglais et de leurs traductions en grec. Ainsi, même si le volume des textes originaux grecs ne permettait pas une analyse complète et bien fondée, les éventuelles différences des résultats des deux étapes auraient pu être imputables à leur absence. Cette distinction a pourtant vite été abandonnée. D'une part, parce que les différences entre les résultats obtenus n'étaient pas importantes et, d'autre part, parce que ces divergences concernaient plutôt des résultats non pertinents dans le cas du corpus contenant seulement les textes originaux en anglais[201].

1.2. Première étape de prétraitement du corpus d'apprentissage

1.2.1. Etiquetage morphosyntaxique et lemmatisation

Le corpus d'apprentissage que nous avons utilisé avait déjà subi certaines étapes de prétraitement. Ces étapes seront décrites comme la **première phase** de prétraitement du corpus.

[201] Des exemples précis de ces résultats seront présentés dans le paragraphe 3 du chapitre 6.

Plus concrètement, les deux parties du corpus (anglaise et grecque) avaient été **étiquetées morphosyntaxiquement** et **lemmatisées**. L'étiquetage morphosyntaxique consiste à attribuer une étiquette de partie du discours à chaque mot d'un texte, tandis que le processus de lemmatisation consiste à associer à chaque occurrence d'un mot sa forme canonique (lemme). L'analyse morphosyntaxique et la lemmatisation des textes anglais du corpus ont été effectuées à l'aide de l'étiqueteur **TreeTagger** (Schmid, 1994) en utilisant le jeu d'étiquettes du **Penn-TreeBank** (Marcus *et al*., 1993). Le TreeTagger est un étiqueteur probabiliste qui évalue la probabilité de séquences de mots étiquetés (c'est-à-dire, la probabilité de transition entre un ensemble d'étiquettes) à l'aide d'un arbre de décisions. L'annotation des textes grecs a, quant à elle, été effectuée avec l'étiqueteur des parties du discours de l'Institut pour le Traitement du Langage et de la Parole[202] (ILSP). L'architecture de cet étiqueteur est similaire à celle de l'étiqueteur à base de transformations de Brill (1995), auquel quelques modifications ont été apportées, permettant un traitement plus efficace des particularités du grec (Papageorgiou *et al*., 2000).

Le jeu d'étiquettes utilisé pour le grec est conforme au jeu d'étiquettes **PAROLE** (Labropoulou *et al*., 1996) et comprend 584 étiquettes de parties du discours différentes. La taille du jeu d'étiquettes utilisé s'explique par la grande quantité d'informations qu'il vise à capter. Les informations encodées pour les noms, par exemple, concernent la partie du discours (nom), le type de partie du discours (nom commun ou propre), le genre (masculin, féminin, neutre), le nombre (singulier, pluriel) et le cas (nominatif, génitif, accusatif, datif, vocatif)[203]. Des traits similaires sont encodés pour les adjectifs et les articles. Un plus grand nombre de traits est encodé pour les pronoms, tandis que les étiquettes les plus longues correspondent aux verbes[204]. Les résultats d'une évaluation de la performance de l'étiqueteur, rapportés dans Papageorgiou *et al*. (*ibid*.), varient selon la quantité d'informations encodées prise en compte. Ainsi, le taux

[202] 'Institute for Language and Speech Processing' (ILSP), Athènes, Grèce.

[203] L'étiquette 'NoCmMaSgNm', par exemple, encode les informations suivantes : No(un), C(o)m(mon), Ma(sculine), S(in)g(ular) et N(o)m(inative).

[204] Ces étiquettes contiennent les traits suivants : type de partie du discours, perfectivité, temps, aspect, voix, nombre, genre et cas.

d'erreur[205] de l'étiqueteur est très faible (4,23%) lorsqu'il s'agit de la catégorie de base mais augmente lorsqu'il s'agit du genre (6,26%). La considération des traits verbaux n'affecte pas beaucoup la performance (6,92%), contrairement à l'ajout de traits d'accord (cas et nombre) qui provoque une augmentation importante du taux d'erreur (10,57%). Même si le grec est une langue hautement flexionnelle, un fort degré d'ambiguïté existe entre formes flexionnelles de parties du discours différentes et même entre formes flexionnelles au sein du même paradigme morphologique. Le premier type d'ambiguïté provoque des erreurs quant à la catégorie de base, tandis que le deuxième type modifie la performance de l'étiqueteur concernant le genre et les traits verbaux, ainsi que l'accord.

Les informations sur les lemmes sont extraites du lexique morphologique de l'ILSP pour le grec. Chaque occurrence d'un mot du corpus est ramenée à sa forme canonique qui lui est ensuite attribuée. Les fichiers contenant les informations d'étiquetage morphosyntaxique et de la lemmatisation sont en format XCES[206].

1.2.2. Alignement phrastique

Le corpus parallèle INTERA est un **bitexte** (Harris, 1988a,b ; Isabelle, 1992), aligné au niveau des phrases[207]. Nous parlons de bitexte dans le cas d'un corpus parallèle où des parties des textes de la LS sont appariées avec des parties des textes de la LC. L'appariement des textes, appelé aussi **alignement**, est un processus qui prend en entrée un corpus parallèle (c'est-à-dire un corpus qui contient des textes originaux dans une langue et leurs traductions dans une autre) et donne à la sortie des appariements entre les deux textes mettant en correspondance les régions textuelles qui sont des traductions l'une de l'autre (Langlais et El-Bèze, 1997). Autrement dit, soit un texte et sa traduction, un

[205] Le taux d'erreur est calculé sur la base du nombre de mots auxquels une étiquette de partie du discours est attribuée. Ainsi, la ponctuation, les chiffres, les dates, les délimiteurs de phrases, etc. reconnus par le tokeniseur, qui effectue aussi la segmentation en phrases, ne sont pas pris en compte lors de cette évaluation.

[206] Corpus Encoding Standard for XML.

[207] Le terme « bitexte » (ou « bi-texte ») est souvent employé dans la littérature pour décrire un corpus parallèle qui contient des versions originales des textes et leurs traductions dans une autre langue. Nous utiliserons ici ce terme pour parler de **corpus bilingues alignés**.

alignement est une segmentation des deux textes telle que le *n*ième segment d'un texte soit la traduction du *n*ième segment de l'autre. L'opération d'alignement consiste à extraire un sous-ensemble du produit cartésien des ensembles de segments source et cible. Les régions textuelles mises en correspondance lors de l'alignement peuvent concerner des unités textuelles plus ou moins grandes, allant du document jusqu'au mot, voire au caractère[208].

Les méthodes proposées pour l'alignement au niveau des phrases se basent souvent sur des propriétés de surface des textes, qui concernent des proportionnalités de la longueur des segments appariés, calculées en termes de mots (Brown *et al*., 1991a) ou en termes de caractères (Gale et Church, 1991, 1993). Ces méthodes reposent sur l'hypothèse qu'il existe une corrélation forte entre les longueurs des segments qui sont traduction l'un de l'autre. D'après cette hypothèse, des phrases longues dans une langue ont tendance à être traduites par des phrases plus longues dans une autre, tandis que des phrases plus courtes sont souvent traduites par des phrases courtes. Ces méthodes s'appuient donc sur les caractéristiques formelles des textes et sont caractérisées comme statistiques.

D'autres méthodes font usage d'informations sur le contenu des régions à aligner. Dans le cas de l'alignement phrastique, ces informations concernent l'appariement des unités lexicales composant les phrases. Le modèle de Kay et Röscheisen (1988, 1993), par exemple, repose sur une relation entre alignements de mots et alignements de phrases, basée sur l'observation que si deux phrases contiennent une paire de mots alignés, elles doivent être, elles aussi, alignées. Dans ce modèle, un alignement lexical partiel sert donc à améliorer l'alignement au niveau des phrases. L'idée de Kay et Röscheisen, qui consiste à faire reposer l'appariement des phrases sur l'appariement des mots, est reprise dans la méthode de Debili et Sammouda (1992). Ces derniers soutiennent que, pour obtenir un appariement fin des mots, il faut apparier les phrases, et que pour apparier les phrases, un appariement grossier des mots serait nécessaire. La comparaison de deux phrases, dans ce modèle, repose sur l'appariement des mots qui les composent (à l'aide d'un dictionnaire de transfert de mots simples) ;

[208] Les unités intermédiaires pouvent être le chapitre, la division, le paragraphe, la phrase, la proposition ou le terme.

plus cet appariement est dense, les mots appariés longs et leur séquentialité respectée, et plus les phrases sont proches.

L'amélioration de l'alignement statistique passe donc souvent par le recours à une petite quantité d'informations linguistiques. La méthode de Simard *et al*. (1992) utilise également de telles informations pour surmonter les faiblesses de la méthode de Gale et Church. Plus précisément, cette méthode exploite les informations de **cognates**, qui sont des paires d'occurrences de mots de langues différentes partageant des propriétés phonologiques, orthographiques voire sémantiques, et qui sont probablement des traductions mutuelles. La méthode proposée par Papageorgiou *et al*. (1994) exploite, quant à elle, des informations linguistiques de surface combinées à des informations sur la charge sémantique d'une phrase, exprimée par les motifs d'étiquettes de parties du discours des mots de contenu qui y apparaissent. Une connexion entre deux unités textuelles est établie si la charge sémantique d'une unité est proche de la charge sémantique de l'autre.

La méthode utilisée pour l'alignement phrastique de notre corpus d'apprentissage repose sur un ensemble de **points d'ancrage** initialement définis entre les textes des deux langues, et qui consistent en des correspondances de mots ou de séquences de mots (Triantafyllou *et al*., 2000)[209]. Les points d'ancrage repérés sont utilisés pour établir des correspondances entre les phrases des textes parallèles. L'alignement phrastique se base sur le modèle de Gale et Church (1991, 1993), qui exploite des informations sur la longueur des phrases en nombre de caractères. L'alignement optimal des phrases s'obtient en faisant appel à une technique de programmation dynamique.

Les phrases alignées du corpus sont regroupées au sein d'**unités de traduction**. Une unité de traduction peut contenir de 0 à 2 phrases par langue. Par exemple, un alignement *2:1* met en correspondance deux phrases du texte de la LS avec une phrase du texte de la LC, au sein d'une unité. Un alignement *1:0* indique un cas d'omission, lorsque la phrase du texte de la LS n'a pas de correspondance dans le texte de la LC, tandis qu'un alignement *0:1* indique un ajout, c'est-à-dire le fait qu'une phrase ait été ajoutée dans la traduction sans

[209] Le processus de repérage de correspondances de longueur arbitraire entre séquences de mots parallèles se base sur l'algorithme de Kitamura et Matsumoto (1997).

qu'une phrase correspondante n'existe dans le texte original. La correspondance de type *2:2* permet de capter les correspondances croisées, c'est-à-dire les cas où l'ordre de 2 phrases dans le texte de la LS est renversé dans la LC. Il faut néanmoins remarquer que la plupart des correspondances établies entre les phrases sont de type *1:1*. Ce type de correspondance est illustré par la figure 2.

<seg id="seg.EN.125"> The Council may, acting unanimously on a proposal from the Commission, decide to shorten or terminate the transitional period indicated in the first paragraph. </seg>
<seg id="seg.EL.125"> Το Συμβούλιο μπορεί, αποφασίζοντας ομόφωνα βάσει προτάσεως της Επιτροπής, να αποφασίσει τη συντόμευση ή λήξη της μεταβατικής περιόδου που αναφέρεται στην πρώτη παράγραφο. </seg>

Figure 2. Unité de traduction regroupant une phrase par langue

Nous parlerons dorénavant de **segments** de la LS et de la LC mis en correspondance au sein d'une unité de traduction et non de phrases. Nous choisissons d'utiliser ce terme en raison du nombre variable de phrases pouvant être incluses dans un segment. Les informations d'alignement des phrases sont encodées dans les fichiers XCES qui contiennent les informations d'étiquetage morphosyntaxique et de lemmatisation, au moyen de numéros (identifiants) attribués aux segments des deux langues.

Les unités de traduction définies lors de l'alignement phrastique nous serviront à délimiter le **contexte** lexical (ou la **fenêtre textuelle**) nécessaire à l'application des méthodes contextuelles que nous proposons sur les données du corpus d'apprentissage.

1.3. Deuxième étape de prétraitement du corpus d'apprentissage

1.3.1. Diagramme de flux de données

Les étapes de prétraitement décrites jusqu'ici (première phase) ont été effectuées au sein de l'ILSP avant le commencement de ce travail. En revanche, la deuxième phase de prétraitement du corpus concerne les étapes qui ont été menées dans le cadre de ce travail de thèse. Ces étapes sont illustrées dans le

diagramme de flux de données inclus en Annexe A1. Ce diagramme global est une représentation graphique du traitement des données lors des trois principales phases de ce travail : le prétraitement, l'apprentissage et l'évaluation. Il permet de structurer et de visualiser le traitement effectué et aide à préciser et à décrire les transformations qui s'opèrent, au sein du système, sur les données d'entrée pour générer des données de sortie.

Dans ce diagramme, les **rectangles** représentent des **entités externes**, qui correspondent à des sources ou à des destinations de données. Les **ellipses** représentent des **processus** qui reçoivent des données en entrée, les traitent et les fournissent en sortie, tandis que les flèches représentent le flux de données. La partie du diagramme qui décrit les étapes de prétraitement du corpus d'apprentissage est reprise dans la figure 2.

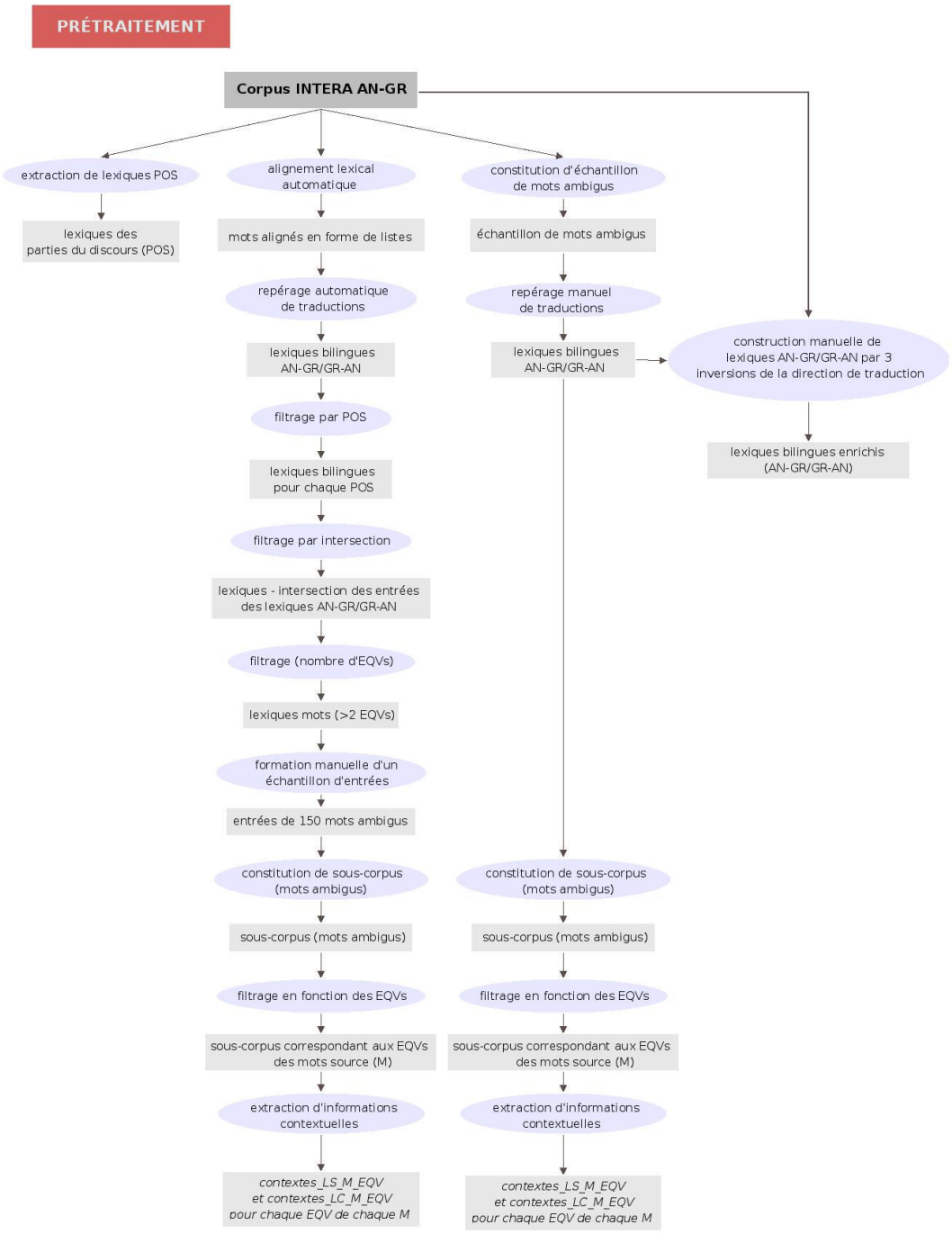


Figure 2. Description du processus de prétraitement

Les données d'entrée du corpus INTERA (AN-GR) sont des fichiers de format XCES, qui contiennent les informations ajoutées lors des étapes de prétraitement décrites dans les paragraphes précédents. A ce stade, le corpus est donc lemmatisé, morphosyntaxiquement étiqueté et aligné au niveau des phrases.

1.3.2. Alignement lexical

La première étape de la deuxième phase de prétraitement consiste à aligner le corpus au niveau des mots. L'objectif de l'alignement lexical est d'identifier des liens de traduction entre mots (ou unités lexicales complexes) des segments source et cible (Brown *et al*., 1993 ; Vogel *et al*., 1996 ; Ahrenberg *et al*., 1998 ; Och et Ney, 2000 ; Tiedemann, 2004). L'alignement lexical s'appuie généralement sur le résultat de l'alignement des phrases et, dans ce cas, le but est de créer des correspondances entre les mots contenus dans des phrases déjà appariées. Il a été justement démontré que l'exploitation de l'appariement des phrases exerce une influence bénéfique sur les résultats de l'alignement lexical (Debili et Sammouda, 1992) : s'appuyer sur l'alignement des phrases permet en effet de réduire la combinatoire en délimitant les champs d'investigation, c'est-à-dire les parcelles de texte dans lesquelles il est possible de trouver l'expression correspondante recherchée. Les expressions mises en correspondance lors de l'alignement lexical sont internes aux phrases et la recherche se limite ainsi à l'intérieur des phrases. Les possibilités de choix étant réduites, les probabilités d'erreurs se voient à leur tour minimisées. La recherche de correspondances dans une paire de phrases fait aussi intervenir les contextes des expressions à apparier ; l'appariement des éléments situés à proximité aide à prendre une décision, ce qui peut également susciter une diminution des erreurs.

La méthode utilisée pour l'alignement lexical de notre corpus prend en entrée des textes alignés au niveau des phrases et utilise deux modèles IBM-2 (Brown *et al*., 1993), un modèle direct (de la LS vers la LC) et un modèle inversé (de la LC vers la LS), qui permettent de prendre en compte le principe de

compositionnalité (Simard et Langlais, 2003)[210]. Le résultat de l'alignement consiste en des correspondances au niveau des mots des phrases alignées.

L'alignement lexical a été effectué deux fois : la première étape traitait les **occurrences** des mots du corpus tandis que, pendant la deuxième étape, ce sont les **formes** auxquelles les occurrences étaient ramenées lors de la lemmatisation qui ont été alignées. L'alignement des formes a été rendu possible grâce aux informations de lemmatisation disponibles au sein des fichiers 'xml' du corpus d'apprentissage. L'exploitation des informations de lemmatisation situe l'alignement lexical à un niveau d'abstraction supérieur (formes de base) à celui des formes de surface (occurrences).

Les résultats obtenus en ce qui concerne les formes canoniques sont de qualité nettement supérieure à ceux traitant les occurrences. La lemmatisation a effectivement un impact bénéfique sur ce type de traitement, surtout lorsqu'il s'agit de langues à morphologie riche (comme c'est le cas du grec). L'effet positif de la prise en compte d'informations linguistiques de ce type sur la TA statistique a été démontré par Nießen et Ney (2000, 2004). Les correspondances entre les mots de la LS et de la LC étant acquises à partir de corpus bilingues sur la base de modèles d'alignement n'utilisant pas (ou très peu) d'informations linguistiques, il est démontré que la TA statistique peut profiter de l'introduction explicite d'une petite quantité de connaissances sur les langues considérées, telles que les informations de lemmatisation ou les informations morphosyntaxiques.

La prise en compte de ces connaissances, d'une part, améliore la qualité de la traduction et, d'autre part, réduit le volume de texte nécessaire à l'entraînement des paramètres des modèles, en optimisant l'exploitation des données bilingues disponibles et en réduisant par là le phénomène de la dispersion des données. Un autre effet de la prise en compte de ces informations est l'amélioration de la couverture lexicale, dans la mesure où ces informations permettent la dérivation de la traduction d'occurrences non connues des mots[211].

[210] L'alignement automatique du corpus au niveau des mots a été effectué par Philippe Langlais.

[211] D'après Nießen et Ney (*ibid.*), les formes fléchies des mots dans la langue d'entrée contiennent souvent des informations qui ne sont pas pertinentes pour la traduction, surtout lorsque la traduction se fait d'une langue hautement flexionnelle dans une autre qui l'est moins. Dans ce cas, des formes fléchies distinctes des mots peuvent être traduites par la même forme dans la LC. Les auteurs proposent la création de classes d'équivalences de formes lexicales, qui ont tendance à être traduites par le même mot dans la LC. Ces classes sont construites en éliminant les informations qui ne sont pas pertinentes pour la traduction.

Il est aussi intéressant de noter que la lemmatisation d'une seule des deux parties du corpus parallèle est considérée comme ayant un effet positif sur la performance de l'alignement statistique des mots (Ozdowska, 2006). Un processus similaire à la lemmatisation est proposé par Fraser et Marcu (2006), relevant d'un ensemble de processus visant à améliorer la qualité de l'alignement lexical. Il s'agit de la **troncation** (*stemming*) – processus plus simple que la lemmatisation – qui consiste à conserver un nombre précis des premiers caractères des mots.

La différence observée au niveau de la qualité des résultats obtenus par l'alignement des formes et des occurrences des mots, justifie le fait que nous utiliserons par la suite les résultats de l'alignement des formes. Dorénavant, lorsque nous parlerons de résultats d'alignement automatique, nous ferons référence aux résultats concernant les formes lexicales.

1.3.3. Création de lexiques bilingues

Les résultats fournis par l'aligneur mettent en évidence les correspondances entre les formes des mots des phrases alignées du corpus. A partir des résultats obtenus par l'alignement lexical, nous avons créé deux **tableaux d'associations** :

1. un tableau où les mots anglais sont mis en correspondance avec leurs traductions potentielles en grec, repérées lors de l'alignement des mots dans l'ensemble du corpus

2. un second tableau pour l'autre direction, où les mots grecs sont mis en correspondance avec les mots anglais, auxquels ils ont été liés lors du processus d'alignement.

Ces deux tableaux, qui contiennent des correspondances au niveau des formes, peuvent être considérés comme des **lexiques bilingues**. L'intérêt de traiter les formes lemmatisées des mots est évident à ce niveau[212]. Chaque

[212] L'avantage de prendre en compte les formes lemmatisées des mots apparaît également lors de l'extraction automatique de lexiques bilingues à partir de corpus alignés au niveau des phrases, sans alignement lexical. Cette méthode permet d'obtenir des calculs plus robustes, avantage majeur lors de l'utilisation de modèles statistiques (Gaussier et Langé, 1995).

équivalent fourni pour un mot source au sein de ces lexiques est accompagné d'un score qui montre la probabilité de son association avec le mot source. Les entrées sont ordonnées par rapport à leur score d'association. La forme des entrées de ces lexiques est décrite dans la figure 3.

mot_source1 : *équivalent_1_1* (score_1_1), *équivalent_1_2* (score_1_2), …, *équivalent_1_n* (score_1_n)

mot_source2 : *équivalent_2_1* (score_2_1), *équivalent_2_2* (score_2_2), …, *équivalent_2_m* (score_2_m)

…

Figure 3. Forme des entrées des lexiques bilingues

La longueur des listes d'associations fournies nous a contrainte à définir un **seuil** au-dessous duquel les associations sont ignorées. Ce seuil peut s'appliquer au score attribué à chaque équivalent, qui indique la force de son lien avec le mot source. Idéalement, la définition d'un tel seuil permettrait d'éliminer les équivalents ne présentant pas une forte association (probablement pertinente) avec le mot source. Nous avons pourtant observé qu'un seuil de ce type peut éliminer des résultats des équivalents pertinents (mais probablement assez rares), qui se situent parfois relativement bas dans la liste des résultats ou, du moins, au-dessous de correspondances erronées ayant, pour quelque raison, reçu un score plus élevé. Une entrée où s'observe ce phénomène est celle correspondant au nom *deprivation*.

deprivation : στερητικός (0,1375) καταδικάζω (0,07845) τρίτος (0,07478) συνεπάγομαι (0,06505) τούτος (0,05881) στέρηση (0,05256) μέτρο (0,04257) υπήκοος (0,03671) αδίκημα (0,03606) ποινή (0,03572) …

Dans cette liste d'équivalents, l'équivalent correct *στέρηση* se situe à la sixième place, avec un score de 0,05256. Le seuil que nous avons finalement retenu détermine le nombre maximal d'associations pour chaque mot. Par souci de préserver, dans les résultats, le maximum d'équivalents pertinents possible pour un mot source, et ce, relativement au nombre total d'équivalents pertinents pour ce mot dans le corpus (ou, mieux encore, pour conserver la totalité des équivalents pertinents), le seuil a été fixé à 15. Ce souci de préserver le maximum

de bons équivalents possible, pourrait s'interpréter comme l'exigence d'un bon **rappel**. L'adoption de ce seuil élevé a eu un effet bénéfique dans le cas de mots ayant plusieurs équivalents dans le corpus. L'effet négatif consiste en ce que ce seuil a laissé pénétrer beaucoup de bruit dans les résultats. Cet effet s'intensifie dans le cas de mots ayant un petit nombre d'équivalents, où le reste des associations proposées constituent du bruit. Par exemple, dans le cas du mot *integrity*, pour lequel un seul équivalent correct est trouvé (*ακεραιότητα*), les autres associations proposées (14/15) sont toutes erronées :

> *integrity* : ακεραιότητα (0,6186) τάξη (0,0704) διαφυλάγω (0,03246) δε (0,02442) ιδίως (0,01949) προβλεπόμενος (0,01868) αποβλέπω (0,01862) ο (0,01705) ενώνω (0,01526) επίσης (0,009939) δικαίωμα (0,009308)

Afin d'éliminer le bruit des résultats, nous leur avons appliqué un ensemble de filtres que nous allons décrire.

1.3.4. Filtrage par partie du discours

Le premier filtrage vise à trier les équivalents fournis au sein des lexiques bilingues en fonction de leur partie du discours. Pour que ce filtrage soit possible, des **lexiques des parties du discours** doivent être disponibles. Ces lexiques ont été construits à partir du corpus d'apprentissage, qui est, nous le rappelons, morphosyntaxiquement étiqueté. Trois lexiques ont ainsi été extraits à propos de catégories différentes : les noms, les verbes et les adjectifs. Si un mot est étiqueté dans le corpus d'apprentissage par une de ces parties du discours, il est automatiquement intégré dans le lexique correspondant. Par exemple, si un mot est toujours étiqueté comme « Nom » au sein du corpus, il est inclus dans le lexique correspondant aux noms. Dans les cas d'ambiguïté catégorielle – où un mot appartient à plusieurs catégories grammaticales et est, par conséquent, étiqueté par des parties du discours différentes dans le corpus d'apprentissage – le mot est inclus dans le lexique de chaque partie du discours correspondant.

Les lexiques élaborés de cette manière sont utilisés pour éliminer de la liste des équivalents obtenus pour un mot source d'une catégorie grammaticale précise, ceux de catégorie grammaticale différente. Si, par exemple, le mot source

est un nom, seuls ses équivalents noms sont retenus[213]. Dans le cas du nom *integrity* cité plus haut, le filtrage par partie du discours permet de garder dans les résultats les associations suivantes :

integrity : ακεραιότητα (0,6186) τάξη (0,0704) δικαίωμα (0,009308) σεβασμός (0,008521)

Ce filtrage diminue donc le bruit de manière importante. Cette hypothèse de correspondance entre unités lexicales en relation de traduction au niveau de la catégorie grammaticale est sous-jacente à un ensemble de travaux qui visent l'étiquetage des mots d'une langue par transfert d'annotations morphosyntaxiques fournies dans une autre langue (Yarowsky *et al*., 2001 ; Borin, 2002 ; Ozdowska, 2006). Le transfert s'appuie précisément sur un alignement au niveau des mots entre les deux langues[214].

D'un point de vue linguistique, cette hypothèse n'est pas considérée comme valide pour toutes les paires de langues ni pour toutes les parties du discours. La validité de l'hypothèse, et la réussite qui s'en suit des systèmes de transfert de ces informations d'une langue à une autre, dépendent de la relation qu'entretiennent les langues concernées[215]. L'hypothèse s'applique plutôt à des langues génétiquement très proches ou à des langues ayant été en contact pendant longtemps (Borin, *ibid*.). Mais, même entre langues proches, il est possible que les catégories grammaticales ne présentent pas toutes la même tendance à l'invariabilité lors de la traduction. Selon Borin (*ibid*.), dans les cas de non

[213] Ce filtrage concerne les lexiques bilingues générés à partir des résultats de l'alignement des mots. Il est donc effectué hors contexte, ce qui peut poser des problèmes dans les cas d'ambiguïté catégorielle. Par exemple, si un mot est utilisé comme nom ou comme verbe au sein du corpus d'apprentissage (disons, le mot *plant*), celui-ci est inclus dans les deux lexiques, des noms et des verbes. Lors du filtrage, si le mot source grec est un nom, le mot *plant* sera retenu comme équivalent du mot source dans la mesure où il appartient au lexique correspondant à cette partie du discours. Il pourrait pourtant arriver que *plant* soit utilisé comme un verbe lorsqu'il traduit le nom en question dans le corpus et qu'il soit donc préservé de façon erronée. Ce problème ne peut être résolu dans un travail hors contexte.

[214] Ce processus permet la création de corpus morphosyntaxiquement étiquetés dans des langues ne disposant pas de telles ressources ou d'outils appropriés pour leur création (c'est-à-dire d'étiqueteurs morphosyntaxiques). Pour que cet étiquetage ait lieu, il suffit qu'un corpus parallèle, un étiqueteur morphosyntaxique (pour la LS) et un outil d'alignement soient disponibles. Borin (*ibid*.) souligne pourtant que l'étiquetage de la LC obtenu de cette manière est « partiel », c'est-à-dire que l'étiquetage initial de granularité grossière pourrait, par la suite, être raffiné.

[215] L'anglais et le grec appartiennent à des branches différentes de la famille des langues indo-européennes : l'anglais est une langue germanique, tandis que le grec fait partie de la branche grecque des langues indo-européennes.

correspondance directe, il serait néanmoins possible de déterminer les conditions qui la provoquent, ou de repérer des régularités servant à la formulation de règles de correspondance de parties du discours pour la traduction entre une paire de langues. La formulation de telles règles rendrait possible, même dans les cas de non correspondance, le transfert d'étiquettes de parties du discours de la LS à la LC, par des liens établis par un algorithme d'alignement des mots.

Même si la possibilité de traduction d'un mot source par un mot de catégorie différente dans une autre langue ne peut être exclue (ou, mieux encore, même si cette possibilité est très forte), l'utilité d'un filtrage des correspondances de traduction fournies dans un lexique bilingue en fonction de leur partie du discours ne peut être estimée qu'en fonction du but visé et du cadre dans lequel le filtrage se situe. Ce type de filtrage est bien adapté aux exigences de l'étude menée dans le cadre de notre travail : cette étude concerne des correspondances lexicales entre mots de deux langues de la même catégorie, à savoir des noms de la LS et leurs équivalents noms dans la LC. Dans la mesure où il s'agit de repérer des correspondances sémantiques, nous trouvons pertinente la remarque faite par Bentivogli *et al*. (2004), dans un cadre de transfert inter-langue d'annotations sémantiques : si un mot source est traduit par un équivalent de catégorie grammaticale différente, l'équivalent en question n'est pas considéré comme un synonyme inter-langue du mot source. Il s'agit donc d'un cas où le transfert de l'étiquette sémantique du mot source à son équivalent serait erroné. Ignorer les équivalents de catégorie différente revient donc ici à ne préserver que les équivalents de traduction qui pourraient être considérés comme des **synonymes inter-langues** du mot source.

Le processus de filtrage des lexiques bilingues décrit ci-dessus fournit **deux lexiques** distincts (un pour chaque direction : anglais-grec et grec-anglais) pour **chacune des catégories grammaticales** retenues (noms, verbes et adjectifs)[216]. Ce filtrage a sensiblement diminué le bruit apparaissant dans les listes d'associations, en ne retenant, d'une part, que les équivalents de la même catégorie grammaticale que le mot source et en éliminant, d'autre part, les mots fonctionnels (articles et prépositions) trouvés dans les listes d'associations, et qui

[216] La création de lexiques pour les verbes et les adjectifs rend possible l'application des méthodes d'acquisition de sens et de désambiguïsation à des mots d'autres parties du discours.

concernaient de nombreuses fausses associations fournies pour les mots de contenu. Ce traitement permet donc une augmentation de la précision au sein des lexiques. Le filtrage par partie de discours a eu pourtant un effet négatif sur le contenu des lexiques, à savoir la diminution du rappel, c'est-à-dire l'élimination d'associations pertinentes due, dans ce cas, à des erreurs au niveau de l'étiquetage morphosyntaxique au sein du corpus d'apprentissage[217].

1.3.5. Filtrage par l'intersection des associations repérées dans les deux directions

Malgré la diminution importante du bruit après le filtrage par partie du discours, des associations erronées pouvaient encore être trouvées au sein des lexiques générés. Pour cette raison, une étape supplémentaire de post-traitement a été menée, qui consistait à ne conserver que les associations se trouvant à l'intersection des lexiques des deux directions correspondants à chaque partie du discours. Autrement dit, seules ont été préservées les associations des mots trouvées dans les résultats des deux alignements, direct et inverse. Les équivalents retenus pour un mot source sont ceux dont la liste d'associations (dans le lexique de l'autre direction) contient le mot en question. Dans l'exemple illustré dans la figure 4, les équivalents retenus pour le mot anglais 'A' sont les mots grecs 'a' et 'c', car 'A' figure dans les listes d'équivalents de ces mots grecs dans la direction grec-anglais. En revanche, 'b' est éliminé de la liste d'équivalents de 'A', parce que l'association 'b-A' ne figure pas dans le lexique de l'autre direction.

AN-GR	GR-AN
A : a, b, c	a : A, D
…	b : G, K, F
	c : A, B, L
	…

Figure 4. Exemple de lexiques d'associations dans les deux directions

[217] Le taux de réussite de l'étiqueteur est assez élevé, mais le résultat de l'étiquetage automatique n'a pas été validé.

Les lexiques résultant de ce filtrage contiennent donc les associations qui se trouvent à l'intersection des lexiques des deux directions obtenus après le filtrage par partie du discours. Ce processus ressemble à la méthode de **traduction inverse** (*back translation*), utilisée par Ivir (1987) pour restreindre l'étude (dans un cadre de linguistique contrastive et de traduction) aux formes de la LC qui peuvent être traduites par la forme originale dans la LS, lorsque l'on inverse la direction de traduction. Ce processus permet, selon Ivir, d'éliminer des différences non relatives, dues aux idiosyncrasies du traducteur ou bien motivées par des stratégies communicatives ou textuelles particulières.

Déterminer l'intersection des alignements constitue une des étapes heuristiques de post-traitement des résultats de l'alignement servant à effectuer une **symétrisation** des modèles d'alignement statistique directionnels (Och *et al*., 1999 ; Och et Ney, 2003 ; Fraser et Marcu, 2006)[218]. L'entraînement s'effectue dans les deux directions de traduction (source-cible et cible-source), ce qui fournit deux alignements pour chaque paire de phrases dans le corpus d'entraînement. La détermination de l'intersection combine les alignements obtenus dans les deux directions. Les éléments situés à l'intersection résultent des deux alignements Viterbi[219] et se révèlent donc très pertinents. Parmi les méthodes proposées par Och et Ney (*ibid*.) pour combiner les résultats des alignements afin d'améliorer leur qualité, il y a l'union, ainsi qu'une méthode plus raffinée, qui consiste à déterminer l'intersection des alignements puis à étendre itérativement cet ensemble, en y ajoutant des alignements qui apparaissent dans l'un des deux alignements initiaux, conformes à un ensemble de conditions. De façon évidente, l'**intersection** des alignements des deux directions donne un alignement qui se caractérise par une précision plus haute et un rappel plus bas que si une seule

[218] La symétrisation des modèles directionnels d'alignement permet l'alignement d'un mot de la LS à plus d'un mot de la LC, ce qui est impossible avec un modèle « baseline ». Un des cas impliquant ce type de correspondance est celui de mots composés en allemand et de leurs équivalents de traduction en anglais, où un mot source unique doit être mis en correspondance avec plusieurs mots cible. L'alignement dans les deux directions et l'application d'un ensemble d'heuristiques permet la prise en compte de ce type de correspondance.

[219] L'alignement Viterbi est l'alignement le plus probable entre deux séquences de mots (Brown *et al*., 1993).

direction était prise en compte. En revanche, leur **union** donne un rappel plus haut et une précision plus basse que chaque alignement séparément[220].

Nous devons noter une différence entre l'étape de détermination de l'intersection effectuée dans les travaux cités ci-dessus et celle que nous avons menée : dans le premier cas, l'intersection concerne les résultats de l'alignement obtenus pour chaque phrase. Le calcul de l'intersection que nous effectuons ici concerne, au contraire, les associations trouvées au sein des lexiques bilingues construits à partir des résultats de l'alignement lexical dans les deux directions. L'effet de cette heuristique est pourtant similaire à celui décrit par Och et Ney (*ibid*.), à savoir une **augmentation de la précision** dans les résultats et une **diminution du rappel**. Dans notre cas, la précision peut être conçue comme le rapport d'équivalents corrects sur l'ensemble d'équivalents proposés, tandis que le rappel exprime le rapport d'équivalents pertinents proposés pour un mot sur le nombre total d'équivalents pertinents pour ce mot dans le corpus. En outre, la définition de l'intersection des alignements contribue, dans notre cas aussi, à une symétrisation des résultats obtenus, c'est-à-dire à l'inclusion des mêmes informations dans les lexiques générés par le calcul de l'intersection. Cette symétrisation est importante pour pouvoir appliquer la méthode des Miroirs Sémantiques à nos données, que nous allons présenter plus loin. Une tentative de symétrisation par détermination de l'union des associations des deux directions a également été entreprise, mais a provoqué une prolifération du bruit trouvé dans les lexiques. C'est pourquoi nous avons abandonné cette approche.

L'application des heuristiques décrites dans ces deux derniers paragraphes a contribué à l'élimination d'une bonne partie du bruit trouvé dans les lexiques initialement construits à partir des résultats de l'alignement automatique, tout en en augmentant la précision. L'amélioration de la qualité des contenus des lexiques de ce point de vue a eu comme corollaire la diminution du rappel, c'est-à-dire l'élimination d'un nombre d'associations pertinentes trouvées dans les lexiques initiaux.

[220] Une précision haute ou un fort rappel, dépend de l'application finale visée. Selon Och et Ney (2003), un rappel haut importe davantage dans des applications comme la TA statistique et, dans ce cas, c'est l'union des alignements qui serait probablement choisi. En revanche, dans des applications lexicographiques, des alignements ayant une grande précision seraient plus intéressants et pourraient être obtenus en déterminant l'intersection des alignements.

1.3.6. Filtrage par nombre d'équivalents

Les informations contenues dans le lexique bilingue anglais-grec, après application de ces filtres, ont été exploitées par la méthode non supervisée d'acquisition de sens proposée dans ce travail[221]. L'une des particularités de cette méthode justifie l'étape suivante de traitement des lexiques obtenus, qui consiste à ne garder que les résultats concernant des mots ayant plus de deux associations. Cette condition est essentielle pour le bon fonctionnement du processus de clustering sémantique, décrit en détail par la suite. Nous expliquerons la raison de cette contrainte en §3.2.1. du chapitre 6.

1.3.7. Formation d'un sous-ensemble des entrées du lexique

Il est important de souligner, au préalable, que la méthode de clustering sémantique proposée est très sensible au bruit. L'existence d'associations erronées au sein des entrées du lexique utilisé (anglais-grec) détériore de manière importante les résultats du clustering et, par conséquent, ceux du processus de repérage de sens. Pour cette raison, nous avons décidé de ne pas prendre en compte les entrées du lexique contenant des associations erronées. Les entrées qui ont été retenues correspondent à 150 noms anglais ayant un nombre élevé d'équivalents (>2) et pour lesquels des associations correctes sont fournies dans le lexique. Contrairement à toutes les autres étapes de filtrage, automatisées, cette étape a été effectuée à la main. Ce traitement manuel aurait pourtant pu être évité si les lexiques générés à partir des résultats de l'alignement lexical contenaient moins de bruit. Néanmoins, dans la mesure où dans ce travail l'alignement lexical ne constitue qu'une étape de pré-traitement, nous avons choisi de ne pas chercher à perfectionner la performance de l'aligneur.

[221] Ces informations n'ont pas été utilisées lors de la première expérience que nous avons menée, qui concerne la méthode monolingue d'acquisition de sens et de désambiguïsation (cf. chapitre 5). Comme nous allons l'expliquer plus loin, cette méthode exploite les résultats du repérage manuel de traductions.

1.3.8. Repérage manuel de traductions

Les étapes de filtrage décrites jusqu'ici s'appliquent aux résultats de l'alignement automatique des mots. Ces résultats nous ont servi à démontrer que les méthodes d'acquisition de sens et de désambiguïsation proposées peuvent être utilisées dans une chaîne de traitement entièrement automatique, lorsque les résultats de l'alignement des mots sont de bonne qualité. Cependant, lors du développement de la méthode nous avons utilisé les résultats d'un processus de repérage manuel de traductions. Ce sont ces données que nous allons également utiliser afin d'illustrer le fonctionnement et les capacités de notre méthode. La focalisation sur un petit ensemble de mots pour lesquels des informations traductionnelles de bonne qualité sont disponibles nous donnera la possibilité d'analyser les points forts de la méthode ainsi que les problèmes qui se présentent et qui ne sont pas liés à la qualité de l'alignement lexical.

Ce choix de mener une étude détaillée sur un petit ensemble de données est en accord avec l'une des raisons pour lesquelles la tâche **lexical sample** (échantillon lexical) a été préférée à la tâche **all words** (tous les mots) lors de la première campagne d'évaluation des systèmes automatiques de désambiguïsation SENSEVAL-1[222]. D'après Kilgarriff et Rosenzweig (2000a), si l'échantillon de mots est bien choisi, la stratégie « lexical sample » est plus informative sur la force et les faiblesses des méthodes de désambiguïsation lexicale que la tâche « all words ». Celle-ci fournirait en effet trop peu d'informations sur les problèmes de mots précis pour permettre une bonne analyse. A propos de notre décision de nous référer à un petit ensemble de

[222] Dans une tâche « all words », le système est évalué par rapport à sa performance de désambiguïsation de tous les mots dans un ensemble de textes. A l'inverse, dans une tâche « lexical sample », le système est évalué sur un ensemble de mots présélectionnés, non connus avant l'évaluation. L'échantillon de mots est construit, les instances des mots sont repérées dans les textes et l'évaluation s'applique seulement à ces instances. L'un des avantages de la tâche « lexical sample » est d'offrir la possibilité aux systèmes désignés pour la tâche « all words » d'y participer, tandis que l'inverse n'est souvent pas possible, à cause du besoin éventuel que peuvent présenter les systèmes en matière de données étiquetées pour l'entraînement. En outre, l'étiquetage sémantique d'occurrences d'un échantillon de mots par des humains est considéré comme plus facile et rapide que l'étiquetage de tous les mots du corpus. Cette tâche ne nécessite pas un dictionnaire complet, mais seulement un ensemble d'entrées dictionnairiques de nombre égal au nombre de mots de l'échantillon.

données, nous trouvons dans Kilgariff (1998b) la citation suivante[223] : « *Lexical sense-tagging is not a well-understood task. When a task is not well-understood, it is wise to find out more about it before doing a lot of it. To find out more about it, it is necessary to look closely at it. There is too much data to look closely at everything. The approved scientific procedure, in such circumstances, is to take a sample* »[224]. Etant d'accord avec ce point de vue, nous avons décidé de procéder ainsi.

Une étape manuelle de **repérage de traductions** (*translation spotting*) a ainsi été menée. Nous ne parlons pas dans ce cas d' « alignement manuel des mots », parce qu'un tel processus présuppose l'appariement de tous les mots inclus dans les phrases des deux langues.

Le repérage de traductions consiste à identifier des équivalents de traduction de mots précis de la LS au sein du corpus parallèle (Véronis et Langlais, 2000 ; Simard, 2003). Ce repérage manuel permet une étude minutieuse des relations entretenues entre les mots des deux langues, l'identification d'équivalents très rares ainsi que la prise en compte de cas tels que les omissions, les ajouts et les reformulations au sein des textes de la LC. L'équivalence de traduction au niveau des mots est néanmoins difficile à déterminer, ce qui explique que le repérage manuel de traductions est une tâche complexe, comme c'est le cas également dans l'alignement manuel (Melamed, 1998). Un tel traitement nécessite beaucoup de temps et c'est pourquoi, le plus souvent, il n'a lieu que sur un petit sous-ensemble de mots du corpus. Nous avons ici choisi un ensemble de mots polysémiques anglais (plus précisément, des noms) ayant un grand nombre d'équivalents de traduction grecs au sein du corpus d'apprentissage.

Les noms source retenus sont les suivants : *plant*, *movement*, *power*, *preparation*, *treatment*, *passage*, *paper*, *communication*, *institution*, *occupation*. La sélection de ces mots s'est faite sur la base de leur **fréquence** dans le corpus d'apprentissage et du **nombre de leurs équivalents de traduction**. Il est bien

[223] Bien que la citation concerne l'étiquetage sémantique, cette procédure est néanmoins fortement liée à la tâche de désambiguïsation lexicale.

[224] L'étiquetage sémantique des mots n'est pas une tâche bien comprise. Lorsqu'une tâche n'est pas bien comprise, il est sage d'en savoir davantage sur elle avant de l'appliquer à une grande échelle. Pour en savoir davantage sur elle, il est nécessaire de l'examiner de près. Il y a trop de données pour tout examiner de près. La procédure scientifique recommandée, dans ces conditions, est d'extraire un échantillon.

connu que la fréquence des mots est proportionnelle au nombre de leurs sens (Zipf, 1945)[225]. Cette hypothèse d'une forte corrélation entre la fréquence des mots et leur complexité sémantique est cruciale pour les travaux d'analyse sémantique. Elle est, par exemple, prise en compte lors de la constitution de l'échantillon de mots utilisé dans la tâche d'échantillon lexical de SENSEVAL (Kilgarriff et Rosenzweig, 2000a ; Kilgarriff, 2002)[226], et permet la considération de mots présentant des degrés différents de polysémie, phénomène reflété dans leur fréquence.

Outre la corrélation entre fréquence des mots et degré de polysémie, qui suffirait à justifier la sélection de mots très fréquents dans un cadre monolingue, nous avons émis une autre hypothèse à propos du nombre d'équivalents de traduction des mots source sélectionnés. Nous avons supposé que les mots de contenu fréquents de la LS (ayant un grand nombre d'occurrences dans le corpus d'apprentissage) et, donc, hautement polysémiques, auraient un grand nombre d'équivalents de traduction au sein du corpus parallèle. Cette hypothèse a été effectivement validée par le repérage des traductions des mots sélectionnés dans notre corpus.

Nous ne pouvons nier qu'il aurait été préférable de travailler sur un échantillon de mots déjà utilisé dans d'autres travaux de désambiguïsation, ce qui aurait permis de comparer les résultats obtenus par la méthode développée ici avec ceux fournis dans d'autres cadres. L'une des possibilités aurait été, par exemple, d'utiliser les noms sélectionnés dans le cadre d'une des campagnes d'évaluation SENSEVAL, repris dans un grand nombre de travaux de désambiguïsation. Cependant, nous n'avons pas pu repérer un nombre d'occurrences satisfaisant de ces mots dans notre corpus d'apprentissage. Ceci s'explique par la nature de notre corpus, constitué, en grande partie, de textes communautaires, alors que les corpus utilisés pour l'entraînement dans le cadre

[225] Plus précisément, le nombre de sens d'un mot est proportionnel à la racine carrée de la fréquence relative du mot dans un corpus (Zipf, 1945 ; 1949 : 75).

[226] L'échantillon a été construit, dans ce cas, par une approche stratifiée, consistant à choisir des mots de manière aléatoire à partir d'ensembles de mots construits sur la base de leur fréquence. D'après Kilgarriff et Rosenzweig (2000a), un échantillon aléatoire simple de mots (sans considération de la fréquence) serait inapproprié car la plupart des mots de l'échantillon, voire la totalité, seraient des mots de basse fréquence en raison de la distribution zipfienne des fréquences. Néanmoins, les mots très fréquents sont considérés comme plus significatifs, puisqu'ils possèdent plus d'occurrences de mots, et qu'ils représentent des cas plus difficiles pour la désambiguïsation.

des exercices SENSEVAL (et SemEval) sont, pour la plupart, constitués de textes de la langue générale[227]. Outre le problème du petit nombre d'occurrences de ces mots, nous avons aussi observé que certains présentaient une polysémie qui n'apparaît pas dans un corpus constitué de textes relevant de domaines tels que ceux trouvés dans notre corpus d'apprentissage[228].

L'extraction manuelle de correspondances de traduction à partir du corpus a été effectuée au niveau des **formes**. Plus précisément, nous avons voulu réunir tous les équivalents possibles du mot source et les mettre en relation avec la forme correspondante (lemme) de ce mot. Les entrées de ce lexique sont construites de la manière suivante : l'ensemble des occurrences d'un mot source, qui correspondent à ses formes fléchies, sont repérées dans le corpus et leurs équivalents dans les traductions. La forme canonique du mot source est ensuite mise en correspondance avec les formes lemmatisées de tous ses équivalents trouvés dans le corpus. Les entrées du lexique bilingue construit de cette manière ont la même forme que celles des lexiques automatiquement générés, décrite dans la figure 4 du paragraphe 1.3.3 (sans les scores).

Le repérage de traductions a été facilité par l'utilisation de l'interface TermOnto[229] (Bourigault *et al*., 2004). Une base de données a été créée contenant l'ensemble des unités de traduction du corpus parallèle, identifiées lors de l'alignement des phrases. Plus précisément, les segments de la LS et de la LC d'une unité de traduction se situent dans deux champs séparés, liés par un identifiant composé du titre du texte d'où ils sont extraits et d'un numéro attribué à l'unité de traduction en question lors de l'alignement, identifiant apparaissant dans un champ distinct. Pour chaque mot polysémique étudié, les unités de traduction du corpus où il apparaît ont été repérées. Ces unités constituent un sous-corpus correspondant au mot polysémique. Le repérage des équivalents de traduction du mot au sein de ce sous-corpus a été effectué à l'aide

[227] Ainsi, le corpus HECTOR, des articles du 'Wall Street Journal' (tirés du Penn Treebank II et du PropBank), des textes du 'British National Corpus' (BNC) et, dans le SemEval, des articles de Wikipedia plus spécialisés sur la programmation et des extraits de biographies (Kilgarriff et Rosenzweig, 2000a ; Kilgarriff, 2002 ; Mihalcea *et al*., 2004 ; Pradhan *et al*., 2007).

[228] Par exemple, des 13 sens (de WordNet) retenus dans SENSEVAL-2 pour le mot *bar*, seul le sens juridique apparaît dans notre corpus d'apprentissage. Les autres sens (« bar » (l'endroit), « comptoir », « savonnette », « mesure » (en musique), etc.) n'y apparaissent pas, ce qui rend impossible la comparaison des résultats obtenus avec ceux d'autres systèmes de désambiguïsation.

[229] Le traitement du corpus d'apprentissage par l'outil SYNTEX a été effectué par Didier Bourigault.

de requêtes SQL ; chaque fois qu'un équivalent de traduction était repéré, les segments de traduction correspondants étaient éliminés ; une requête portait ensuite sur le reste des segments, jusqu'à ce que tous les équivalents aient été trouvés et qu'il ne reste plus de segments de traduction, ou que les segments restant contiennent des cas de traduction non pris en compte, comme les paraphrases. La consultation des résultats de l'alignement automatique a également favorisé, d'une certaine manière, le repérage de ces associations[230].

Il faut souligner qu'au moment du repérage manuel de traductions, nous n'avons pas pris en compte les cas où la traduction des mots (noms) source présente certaines particularités :

a. cas où le nom polysémique est traduit dans la LC par un mot d'une autre catégorie grammaticale (par ex. un adjectif). Citons, par exemple, les cas où le nom de la LS fait partie d'un terme complexe, traduit également dans la LC par un terme complexe (ainsi, pour les termes complexes suivants : *products of* ***plant*** *origin* / *προϊόντα* ***φυτικής*** *προέλευσης* (proionta fytikis proelefsis), ***plant*** *growth* / ***φυτική*** *ανάπτυξη* (fytiki anaptiksi), *plant matter* / ***φυτικό*** *υλικό* (fytiko yliko), *plant species* / ***φυτικό*** *είδος* (fytiko eidos), le nom *plant* est traduit par l'adjectif *φυτικός, -ή, -ό*).

b. cas où le nom polysémique fait partie d'une expression ou d'un terme complexe lexicalisé(e) dans la LC (par ex. ***plant protection*** *product* : ***φυτοπροστατευτικό*** *προϊόν* (fytoprostateftiko proion) ou *προϊόν* ***φυτοπροστασίας*** (proion fytoprostasias), ***plant health*** *legislation* : ***φυτουγειονομική*** *νομοθεσία* (fytoygeionomiki nomothesia), ***plant health*** *control* : ***φυτουγειονομικός*** *έλεγχος* (fytoygeionomikos elenchos), ***plant genetic*** *diversity* : ***φυτογενετική*** *ποικιλότητα* (fytogenetiki poikilotita)) ou par un nom (***milkweed plant*** / ***ασκληπιάδα*** (asklipiada)).

[230] Lorsque nous avons obtenu les résultats de l'alignement automatique, le repérage manuel de traductions avait déjà été effectué pour la plupart des mots polysémiques étudiés.

c. cas où le nom polysémique est traduit par une paraphrase dans la LC[231]

d. cas de correspondance phrastique, où la phrase de la LC a le même sens que la phrase de la LS, sans qu'il n'y ait de correspondances au niveau des mots ou, du moins, de tous les mots des phrases (par ex. *the argolic orchid is a native plant* / *στην περιοχή αυτοφύεται η αργολική ορχιδέα* (stin periochi aytofyetai i argoliki orchidea)[232]).

Dans le paragraphe sur le filtrage des lexiques bilingues automatiquement générés en fonction des parties de discours (§1.3.4.), nous avons présenté les raisons qui expliquent la préservation d'associations entre des mots des deux langues appartenant à la même catégorie.

Dans le deuxième cas décrit ci-dessus, la correspondance traductionnelle établie doit impliquer des unités plus grandes que le mot (dans au moins l'une des deux langues), ce qui est souvent dû aux vides lexicaux au niveau du vocabulaire de l'une des langues. Dans les exemples cités pour ce cas, les concepts lexicalisés dans la LC impliquent plus d'un mot de la LS, ce qui empêche le traitement du mot polysémique (*plant*) seul.

Dans les deux derniers cas (c et d), la correspondance doit aussi être établie à un niveau autre que celui des mots, pouvant même s'appliquer à la phrase entière. L'impossibilité d'établir des correspondances au niveau lexical, dans les cas précités, explique pourquoi ils ont été ignorés. Les trois derniers cas ne peuvent pas être pris en compte lors de l'alignement automatique à cause de la difficulté à les repérer[233]. L'alignement automatique des mots créerait des correspondances entre les parties des expressions (termes complexes, etc.) des deux langues, à moins que ces combinaisons de mots ne soient identifiées, ou que des réglages appropriés pour la mise en correspondance de plus d'un mot par langue ne soient effectués[234].

[231] Mauranen (2002) décrit les cas où des paraphrases ou des additions sont utilisées pour traduire une unité de la LS comme des cas de changement de niveau et elle souligne que cela advient souvent avec des concepts spécifiques à une culture.
[232] Dans cet exemple, le texte original est en grec et la traduction est en anglais.
[233] Dans ce cas, il faudrait que les unités à associer soient d'abord repérées.
[234] Car les algorithmes *baseline* d'alignement des mots permettent à chaque occurrence de la LS ou de la LC d'être connectée à une seule occurrence dans l'autre langue.

Les résultats du repérage manuel de traductions ont été exploités par les deux méthodes d'acquisition de sens, qui seront présentées par la suite : la méthode monolingue, exposée dans le chapitre 5, et la méthode combinant des informations contextuelles et traductionnelles, présentée dans le chapitre 6.

Les résultats de ce repérage peuvent être présentés directement sous la forme de lexiques bilingues. Ces résultats ne contenant bien évidemment pas de bruit, il est donc inutile de procéder aux étapes de filtrage nécessaires pour les résultats de l'alignement automatique. Cette démarche est illustrée par le diagramme de la figure 3, où une flèche lie le carré décrivant les lexiques bilingues construits manuellement avec le rectangle arrondi qui représente le processus de constitution de sous-corpus. Ce processus concerne les résultats du repérage de traduction ainsi que les résultats de l'alignement automatique. La constitution de sous-corpus pour les mots source représente une étape essentielle pour l'application des méthodes d'acquisition de sens qui seront présentées plus loin.

1.3.9. Construction de lexiques bilingues dans les deux directions de traduction

La deuxième étape de repérage manuel de traductions est nécessaire pour appliquer la méthode des Miroirs Sémantiques (Dyvik, 2003, 2005) aux données de notre corpus. L'identification de correspondances de traduction entre les mots des deux langues et la construction de lexiques bilingues dans les deux directions sont en effet nécessaires à l'application de cette méthode aux données d'un corpus parallèle. Néanmoins, pour que les résultats de l'analyse sémantique soient de bonne qualité, les informations fournies au sein des lexiques doivent être caractérisées par une **haute précision** et un **bon rappel**. Comme nous l'avons déjà vu, les filtrages successifs des résultats obtenus par l'alignement automatique des mots en ont augmenté la précision mais en ont aussi diminué le rappel. C'est la raison pour laquelle les informations incluses dans les entrées du thesaurus sémantique créé par la méthode des Miroirs, lorsqu'elle est appliquée à

ces résultats, ne sont pas suffisantes à une description complète de la sémantique des mots.

L'objectif de l'application de la méthode des Miroirs Sémantiques sur notre corpus est de comparer les résultats acquis par cette méthode avec les résultats obtenus par la méthode d'acquisition de sens que nous proposons. L'idée était qu'une telle comparaison pouvait servir à valider les résultats obtenus par notre méthode[235]. Comme nous l'avons dit dans le paragraphe précédent, nous nous sommes appuyée, pour le développement de cette méthode et l'analyse de sa performance, sur les données du repérage manuel de traductions, c'est-à-dire les informations traductionnelles non bruitées. Pour que cette comparaison soit possible, nous avons dû par conséquent limiter l'amplitude de l'application des Miroirs Sémantiques sur nos données, ce qui nous a permis d'obtenir des descriptions de la meilleure qualité possible. Plus précisément, la méthode des Miroirs a été appliquée à un sous-ensemble des mots polysémiques qui nous ont servi d'exemples lors du développement et de l'analyse de notre méthode. Pour cet ensemble de mots, nous avons procédé à un repérage manuel de traductions dans les deux directions[236].

Il est vrai que les questions qui surgissent lors de l'extraction manuelle de correspondances de traduction au niveau des mots, à partir d'un corpus parallèle, sont nombreuses et les réponses pas toujours évidentes. L'adoption d'un ensemble de critères facilite donc cette tâche et assure la cohérence des résultats de l'extraction. Pour notre part, nous avons suivi les principes proposés par Thunes (2004), qui déterminent ce qui peut, et ce qui ne peut pas, être considéré comme une correspondance de traduction. Ces principes sont influencés par la finalité dirigeant l'extraction de correspondances utilisables par la méthode des Miroirs Sémantiques, à savoir la création d'un **semi-treillis sémantique**. Ces principes visent donc le repérage de correspondances de traduction entre les mots des deux langues et ont été conçus pour des mots et des

[235] Nous reviendrons sur ce sujet dans le chapitre 4.

[236] Il faut néanmoins préciser que la méthode des Miroirs a été appliquée à l'ensemble des noms du corpus (contenus dans le lexique correspondant à cette catégorie grammaticale et construit à partir des résultats de l'alignement automatique), mais que les résultats de ce traitement n'ont pas été exploités. Le faible taux de rappel concernant les traductions possibles des mots trouvés dans les données automatiquement obtenues, n'a pas permis à la méthode des Miroirs de donner des résultats de qualité telle qu'elle permette une comparaison fructueuse avec ceux obtenus par notre méthode d'acquisition de sens.

expressions appartenant aux catégories lexicales ouvertes[237]. Les cas litigieux que nous avons rencontrés étaient nombreux et variés – ce qui est logique lorsqu'il s'agit de textes réels. Ils montrent la difficulté à trouver une définition correspondant à toutes les occurrences. Pour ces cas, il a donc fallu prendre des décisions *ad hoc*.

Comme nous l'avons vu (cf. §1.3.8.), les associations extraites manuellement concernent les formes des mots. Au sein des entrées des lexiques créés, la forme d'un mot source est mise en correspondance avec les formes lemmatisées de tous ses équivalents trouvés dans les textes. Le repérage de traductions concerne les équivalents grecs des noms polysémiques anglais qui constituent l'échantillon de mots étudié. Pour appliquer la méthode des Miroirs Sémantiques au même échantillon, nous avons dû inverser la direction de traduction et chercher les correspondants grecs des équivalents anglais, repérés à l'étape précédente, dans l'ensemble du corpus. Au total la direction de traduction a été inversée trois fois, ce qui est, d'après Dyvik, suffisant pour obtenir des résultats fiables par la méthode des Miroirs Sémantiques. Les étapes d'inversion de la direction de traduction sont décrites dans la figure 5.

[237] Thunes (*ibid*.) préfère parler de « correspondant de traduction » dans ce contexte, à la place d' « équivalent de traduction », car parler d'équivalents de traduction requiert une définition des critères qui permettraient de dire si l'équivalence de traduction a été atteinte. Le terme « équivalence de traduction » a une dénotation plus étroite que le terme « correspondance de traduction ».

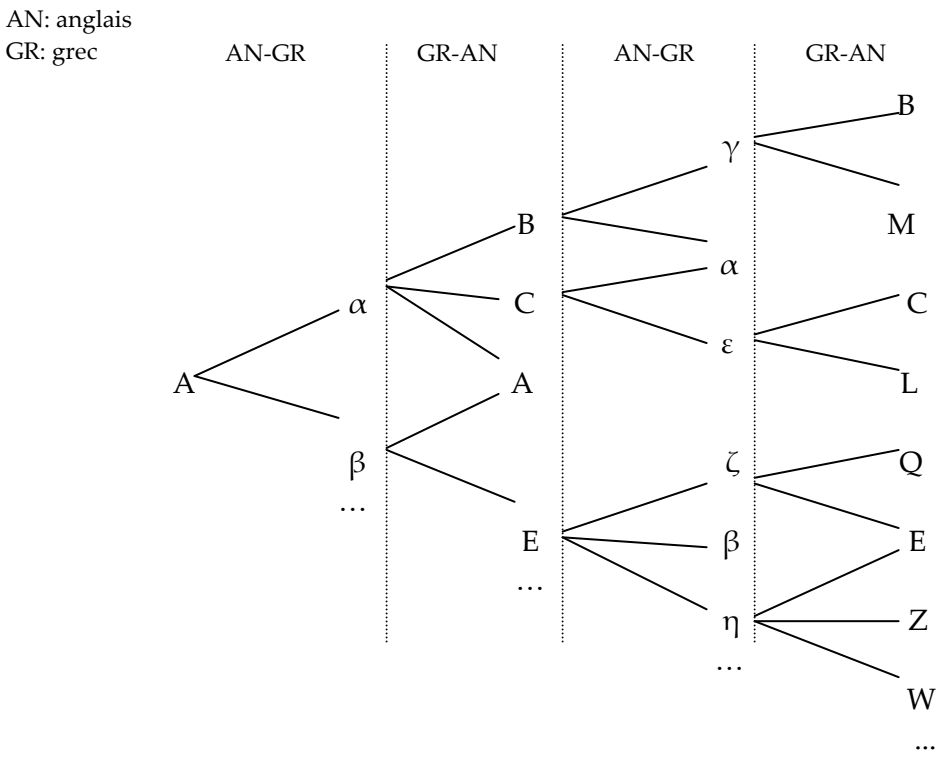


Figure 5. Repérage d'équivalents par inversion de la direction de traduction

A chaque inversion, les mots source correspondent aux équivalents repérés lors de l'inversion précédente. Après avoir repéré les équivalents grecs du mot polysémique étudié, les équivalents anglais de chacun de ces mots grecs sont donc trouvés. La troisième étape consiste à repérer les équivalents grecs des mots anglais extraits lors de l'étape précédente, et dont les équivalents ne se trouvent pas déjà dans le lexique anglais-grec. Finalement, à la quatrième étape, nous inversons encore une fois la direction de traduction, en repérant dans le corpus les équivalents anglais des mots grecs repérés pendant l'étape précédente, et dont les équivalents ne se trouvent pas déjà dans le lexique grec-anglais.

Pour chacun des mots polysémiques étudiés, deux lexiques ont été créés : un contenant les équivalents grecs de tous les mots anglais trouvés après les trois inversions de la direction de traduction, et un autre contenant les équivalents de tous les mots grecs trouvés après ces inversions.

La taille de ces lexiques est d'autant plus grande que le nombre d'inversions de la direction de traduction est grand. Pour le mot *plant*, par exemple, le lexique créé suite aux quatre inversions pour la direction anglais-grec contient 9 entrées

(mots anglais et leurs équivalents grecs) et celui créé pour la direction grec-anglais contient 64 entrées (mots grecs et leurs équivalents anglais). Le très grand nombre de mots grecs s'explique par le fait que la dernière inversion concerne la direction grec-anglais[238]. Bien évidemment, ce processus requiert beaucoup de temps du fait que, non seulement les équivalents de chaque instance du mot polysémique doivent être repérés dans le corpus, mais également les équivalents de ses équivalents, et ainsi de suite.

Dans le temps inévitablement limité que nous avons pu consacrer à cette tâche[239], nous avons obtenu des résultats pour un sous-ensemble de l'échantillon de mots utilisé par notre méthode. Ces mots sont les 5 noms polysémiques anglais suivants : *plant*, *movement*, *treatment*, *occupation* et *power*. Les lexiques créés pour chacun de ces mots ont été finalement fusionnés en créant un lexique global pour la direction anglais-grec et un autre pour la direction grec-anglais. La méthode des Miroirs Sémantiques a été appliquée sur l'ensemble des données, créant un thésaurus comprenant des informations riches sur les mots impliqués. Le lexique global anglais-grec contient 248 entrées (mots anglais et leurs équivalents en grec), tandis que le lexique global grec-anglais contient 661 entrées (mots grecs et leurs équivalents anglais). Deux lexiques ont en outre été créés, contenant seulement les équivalents de traduction qui apparaissent plus d'une fois dans le corpus, c'est-à-dire en éliminant les *hapax* des lexiques globaux[240].

1.3.10. Constitution de sous-corpus

Une fois les lexiques bilingues constitués dans les deux directions, l'étape suivante de prétraitement consiste à répartir le corpus d'apprentissage en un ensemble de sous-corpus. Chacun de ces sous-corpus correspond à un des mots polysémiques étudiés. Comme nous l'avons déjà précisé, dans le cas du repérage

[238] Le nombre de mots anglais aurait donc été beaucoup plus grand si l'on avait inversé la direction de traduction une cinquième fois.

[239] Ce traitement a été réalisé lors de mon séjour de recherche de trois mois à l'Université de Bergen, dans le cadre de la bourse d'accueil Marie Curie pour la Formation de Chercheurs en Début de Carrière.

[240] Les résultats de l'application de la méthode des Miroirs sur les lexiques contenant les instances *hapax* et sur ceux qui ne les contiennent pas, ont été comparés afin d'estimer l'impact des *hapax* sur le résultat de l'analyse sémantique.

manuel, il s'agit de 10 noms polysémiques anglais, tandis que dans le cas des lexiques construits à partir des résultats de l'alignement automatique, nous n'avons retenu que les informations concernant 150 noms anglais ayant un grand nombre d'équivalents, et pour lesquels des associations correctes ont été repérées. Le processus de constitution des sous-corpus exploite le résultat de l'alignement des phrases effectué sur le corpus d'apprentissage. La sortie de l'alignement consiste en des « unités de traduction » comprenant des segments de la LS et de la LC mises en correspondance. Un segment source contient de 0 à 2 phrases de la LS qui se trouve(nt) en relation de traduction avec la phrase ou les phrases contenues dans le segment de la LC. Les segments source et cible qui constituent une unité de traduction ont le même identifiant. Un exemple d'unité de traduction est donné dans la figure 6.

<seg id="seg.AN…">The above provisions do not apply to aid to the transport sector, nor to activities linked to the production, processing or marketing of products listed in Annex I to the EC Treaty with the exception of fisheries products and products derived thereof. **</seg>**

<seg id="seg.GR…">Οι ανωτέρω διατάξεις δεν εφαρμόζονται στις ενισχύσεις στον τομέα των μεταφορών, ούτε σε δραστηριότητες σχετικές με την παραγωγή, επεξεργασία ή εμπορία προϊόντων που απαριθμούνται στο Παράρτημα I της Συνθήκης ΕΚ, εξαιρέσει των προϊόντων αλιείας και των παράγωγων προϊόντων τους. **</seg>**

Figure 6. Exemple d'unité de traduction

Le sous-corpus correspondant à un mot polysémique (*m*) comprend les unités de traduction où ce mot apparaît au sein du segment source. Les différentes instances du mot au sein du sous-corpus peuvent être traduites par des équivalents de traduction différents (*a*, *b*, *c*, etc.) dans les segments de la LC. Ce type de sous-corpus est illustré dans la figure 7.

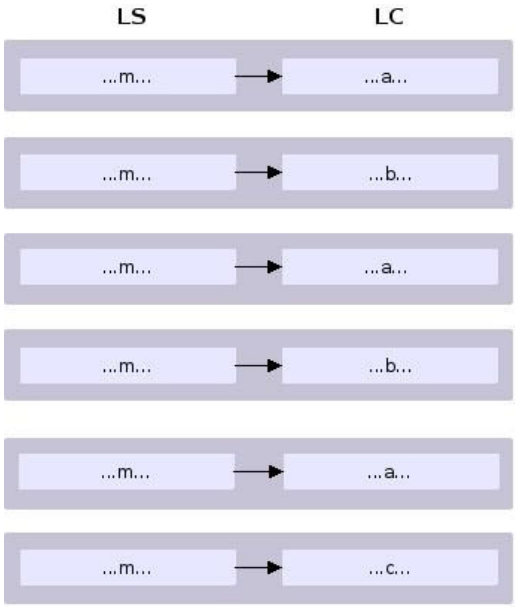


Figure 7. Sous-corpus correspondant à un mot polysémique (*m*) de la LS

1.3.11. Filtrage des sous-corpus en fonction des équivalents

Une fois le sous-corpus correspondant à un mot polysémique constitué, nous procédons à un regroupement des unités de traduction qui y sont incluses, en fonction des équivalents (EQVs) qui traduisent le mot dans les segments de la LC. Lors de cette étape, des ensembles de segments correspondant à chacun des équivalents sont créés du côté de la LS et du côté de la LC (cf. figure 9). Cette figure décrit le filtrage du sous-corpus correspondant au mot polysémique source *m*, traduit dans le corpus par trois équivalents différents : *a*, *b* et *c*. Dans la partie gauche de la figure, se situent les segments de la LS et dans la partie droite, leurs traductions (les segments de la LC) qui apparaissent dans les mêmes unités de traduction, comme cela a été déterminé par le processus d'alignement des phrases.

Ce filtrage se base sur le simple repérage au sein du segment de la LC d'un des équivalents possibles du mot source, identifiés pendant l'étape de repérage manuel de traductions. L'alignement lexical n'est, par conséquent, pas nécessaire pour effectuer ce filtrage. Néanmoins, dans la mesure où il peut arriver que plus d'un des équivalents possibles du mot source apparaissent dans le segment cible,

nous n'avons gardé que les segments contenant une instance d'un seul équivalent et nous avons ignoré les autres[241]. Lorsque plus d'un équivalent apparaissent dans la traduction, il se peut que l'un d'eux traduise un autre mot du segment source (et non le mot polysémique). Etant donné que nous n'utilisons pas les résultats de l'alignement lexical, nous ne pouvons pas savoir quel équivalent traduit effectivement l'instance du mot polysémique. C'est pour cette raison que les unités de traduction dans lesquelles plus d'un équivalent apparaissent dans le segment cible sont ignorées.

Dans la figure 8, les ensembles de segments constitués de cette manière dans les deux langues sont décrits comme « segments_LS_EQV » et « segments_LC_EQV ». L'ensemble « segments_LS_*a* », par exemple, contient les segments de la LS qui comprennent les instances du mot polysémique traduites par l'équivalent *a* dans le corpus[242]. L'ensemble « segments_LC_*a* » contient les segments de la LC correspondant aux segments inclus dans le premier ensemble (qui se trouvent dans la même unité de traduction). Ces segments contiennent, bien évidemment, l'équivalent *a*. Nous procédons de la même manière pour les autres équivalents du mot *m* (*b* et *c*) en constituant les groupes de segments respectifs dans les deux langues.

[241] Nous retrouvons ce type de filtrage dans d'autres travaux comme celui de Vickrey *et al*. (2005), qui vise la prédiction de traduction pour de nouvelles instances de mots ambigus de la LS. Pour chaque instance du mot ambigu dans la phrase source, au sein du corpus d'évaluation utilisé, sa traduction dans la phrase correspondante de la LC est identifiée. Si un seul des équivalents possibles est trouvé, il est remplacé par un « vide » que le module de prédiction de traduction essaie de combler par la suite.

[242] Il faut souligner ici que, dans le cas où deux phrases de la LS sont mises en correspondance avec une phrase de la LC au sein d'une unité de traduction, le mot polysémique est contenu dans l'une des deux phrases. Nous les retenons néanmoins toutes les deux. De même, si une phrase de la LS est mise en correspondance avec deux phrases dans la LC, nous retenons les deux phrases de la LC, et non uniquement celle qui contient l'équivalent en question.

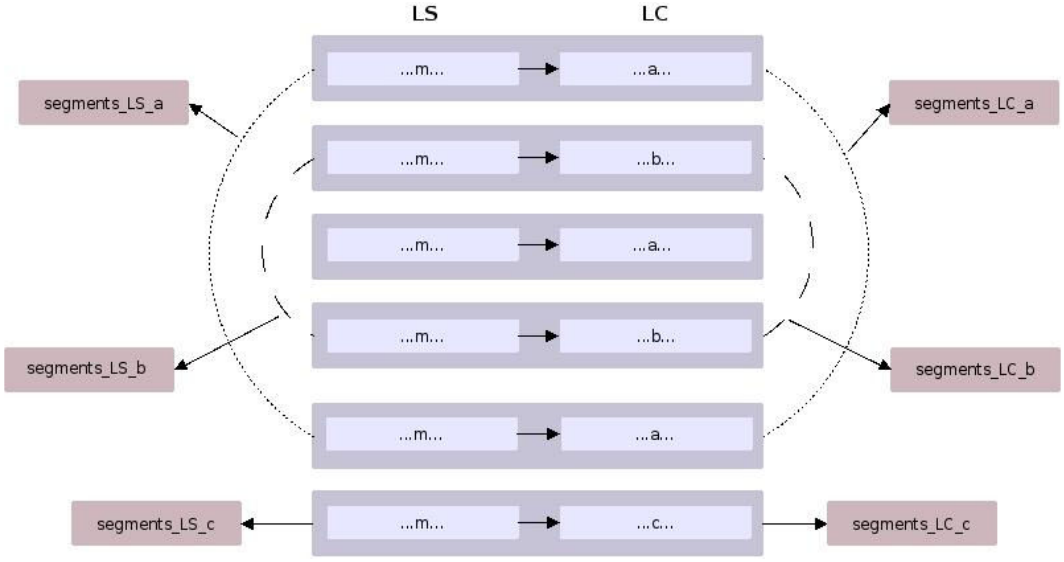


Figure 8. Regroupement des segments des deux langues en fonction des équivalents

La création des sous-corpus correspondant aux mots polysémiques et des ensembles de segments correspondant à leurs équivalents constituent des étapes essentielles pour que les données soient utilisables pendant le processus d'apprentissage des méthodes d'acquisition de sens et de désambiguïsation, présentées par la suite. Comme nous le montrerons plus loin, l'apprentissage peut opérer sur les ensembles de segments des deux langues correspondant aux différents équivalents du mot source (segments_LS_*a* – LS_*b* –_LS *c* etc. / segments_LC_*a* – LC_*b* – LC_*c* etc.). Le processus d'apprentissage sera décrit en détail dans le chapitre 6.

2. Corpus d'évaluation

2.1. Caractéristiques du corpus d'évaluation

Le corpus utilisé pour l'évaluation de la méthode de désambiguïsation et de prédiction de traduction proposée dans ce travail diffère du corpus d'apprentissage. Il s'agit de la partie anglais-grec du corpus parallèle multilingue **EUROPARL** (Koehn, 2003, 2005). Ce corpus est constitué de textes extraits des

actes du Parlement Européen et il est disponible dans onze langues[243]. Le corpus a été traité afin de le rendre conforme à l'utilisation dans la Traduction Automatique statistique. Les textes bruts du corpus ont été obtenus à partir du site Web du Parlement Européen et ont été alignés au niveau du document. Dans la version initiale du corpus, les textes sont également disponibles dans un autre format, alignés au niveau des phrases. L'alignement a été réalisé en utilisant l'algorithme de Gale et Church (1993), basé sur la similarité de la longueur des phrases des deux langues en termes de nombre de mots. Les données alignées au niveau des phrases sont fournies dans des fichiers différents pour chaque langue, de manière à ce que les phrases alignées de deux langues se situent sur la même ligne dans chacun des fichiers correspondants[244].

Nous utilisons ici la version initiale du corpus, la version '1.1', qui contient des données d'avril 1996 jusqu'à décembre 2001. Cette version comprend environ 20 millions de mots par langue, dans 740 000 phrases. Nous avons choisi cette version en raison de la disponibilité de textes alignés au niveau des phrases[245].

2.2. Prétraitement du corpus d'évaluation

2.2.1. Etapes de prétraitement

Avant d'être utilisé pour l'évaluation, le corpus de test a subi, lui aussi, des étapes de prétraitement. La première étape a consisté à filtrer le bruit présent au sein des fichiers grecs, dû à des erreurs de segmentation des mots. Le corpus a été ensuite étiqueté morpho-syntaxiquement et lemmatisé. Pour pouvoir procéder à ces prétraitements, il a d'abord fallu effectuer certaines étapes d'apprentissage, qui nous ont fourni le matériel et l'outil nécessaires (pour la première et la deuxième étape, respectivement). Ce processus est décrit dans la partie du diagramme de flux de données reprise dans la figure 9.

[243] Français, italien, espagnol, portugais, anglais, hollandais, allemand, danois, suédois, grec et finnois.
[244] Par exemple, le contenu de la quatrième ligne de l'un des deux fichiers est aligné avec le contenu de la quatrième ligne de l'autre.
[245] La qualité de l'alignement phrastique est très bonne (Koehn, *ibid*.).

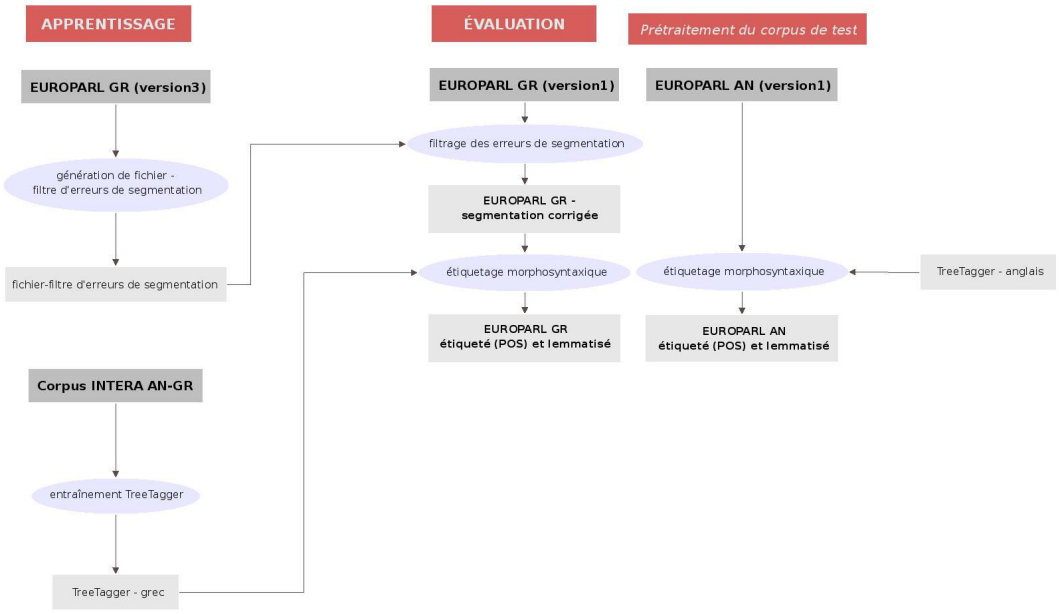


Figure 9. Apprentissage pour le prétraitement du corpus de test

Les étapes finales de prétraitement consistent à élaborer des échantillons de test des mots ambigus puis à les filtrer en fonction des EQVs de traduction de ces mots au sein du corpus de test. Ces étapes sont décrites dans la partie du diagramme de flux de données, reprise dans la figure 10. Nous détaillerons plus précisément chacune de ces étapes dans les paragraphes suivants.

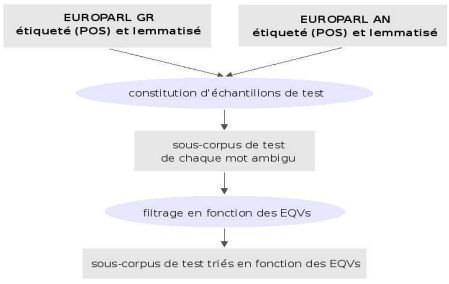


Figure 10. Constitution d'échantillons de test

2.2.2. Correction de la segmentation en mots

La partie grecque de la première version du corpus EUROPARL, que nous avons utilisée, contenait du bruit, dû au découpage erroné des mots

contenant un caractère précis : ' χ'. Ce découpage consistait à inserer une espace avant et après le caractère en question. Ce phénomène aurait un impact sur la qualité de la tokenisation[246], qui influencerait, à son tour, la qualité de l'étiquetage morpho-syntaxique et de la lemmatisation. Pour remédier à ce problème, nous avons construit un filtre à partir de la troisième version du corpus EUROPARL, où ce problème avait été résolu. Ce filtre a alors permis d'éliminer les espaces insérées au sein des mots qui contenaient ce caractère spécifique, dans les textes de la première version. Le filtrage ainsi effectué a permis à la tokenisation, qui a eu lieu lors de l'étiquetage et de la lemmatisation, d'être de bonne qualité.

2.2.3. Etiquetage morpho-syntaxique et lemmatisation

L'apprentissage des méthodes non supervisées d'acquisition de sens et de désambiguïsation que nous proposons opère sur les **formes** de mots de catégories grammaticales précises, trouvés dans les contextes des mots polysémiques étudiés. Ces informations sont aussi utilisées pour la modélisation des correspondances sémantiques inter-langues – entre les mots ambigus et leurs équivalents de traduction – établies lors de l'apprentissage. La désambiguïsation et la prédiction de traduction pour de nouvelles instances des mots polysémiques, repérées dans le corpus de test, sont effectuées en comparant le contexte de ces nouvelles instances aux informations qui représentent les correspondances établies.

La comparaison de ces informations contextuelles à l'**état brut** (c'est-à-dire, l'état dans lequel elles apparaissent dans les textes) aux informations modélisées, qui correspondent aux formes des mots, ne serait pas possible. En outre, étant donné que l'apport des éléments contextuels pour la désambiguïsation est variable, le repérage des éléments contextuels pertinents n'est pas évident en raison du manque d'informations morphosyntaxiques au sein du corpus de test. Par conséquent, pour que la comparaison entre informations des nouveaux contextes et informations modélisées soit possible, il a fallu lemmatiser et étiqueter morphosyntaxiquement les textes du corpus de test.

[246] La tokenisation est l'opération de segmenter un texte en mots.

Pour la lemmatisation et l'étiquetage de la partie anglaise du corpus anglais-grec EUROPARL, nous avons utilisé l'étiqueteur **TreeTagger** (Schmid, 1994) et le fichier de paramètres disponible pour l'anglais. Le jeu d'étiquettes utilisé consiste en un raffinement du jeu d'étiquettes **Penn-Treebank** (Santorini, 1991 ; Marcus *et al*., 1993)[247]. Cependant, un tel fichier de paramètres n'étant pas disponible pour le grec, nous avons dû le créer nous-même. L'entraînement du TreeTagger, nécessaire pour la création de ce fichier, a été effectué sur la partie grecque de notre corpus d'apprentissage. A partir du corpus, nous avons généré :

- un **lexique** : fichier contenant le lexique des formes pleines des mots. Chaque ligne du lexique correspond à une occurrence (forme fléchie) d'un mot et contient l'occurrence en question et une séquence de paires étiquette-lemme. Par exemple :

πηλίκου NoCmNeSgGe πηλίκο
απομίμησης NoCmFeSgGe απομίμηση
διατυπώθηκαν VbMnIdPa03PlXxPePvXx διατυπώνω

- un **fichier de mots de classe ouverte** : fichier contenant la liste des étiquettes de mots de classe ouverte (mots de contenu), qui constituent des étiquettes possibles pour des occurrences de mots inconnues. L'étiqueteur se réfère à ces informations lorsqu'il rencontre des mots inconnus, c'est-à-dire des mots qui ne sont pas contenus dans le lexique.

- un **fichier d'entrée** : fichier contenant des données d'entraînement étiquetées dans un format « un mot par ligne ». Chaque ligne contient donc une occurrence et une étiquette.

Les informations comprises dans ces trois fichiers ont été extraites du corpus d'apprentissage, corpus lemmatisé et morphosyntaxiquement étiqueté. A l'aide de ces fichiers, nous avons créé le fichier de paramètres adapté au grec, qui a été ensuite utilisé pour l'étiquetage et la lemmatisation de la partie grecque du corpus d'évaluation.

Il faut noter ici que l'étiquetage et la lemmatisation du corpus d'apprentissage n'avaient pas été validés à la main. Par conséquent, le fichier de paramètres constitué à partir de ce corpus est « bruité ». L'exploitation des

[247] Ce raffinement permet de distinguer les verbes *be* et *have* des autres verbes.

informations incluses dans ce fichier pour l'annotation du corpus de test provoque ainsi des erreurs au niveau de l'étiquetage morphosyntaxique et de la lemmatisation au sein du corpus de test.

2.2.4. Constitution de sous-corpus d'évaluation

Pour chaque mot polysémique étudié, des sous-corpus ont été constitués à partir du corpus d'évaluation, de la même manière que dans le cas du corpus d'apprentissage (cf. §1.3.10.). Chaque sous-corpus contient les unités de traduction où le mot polysémique apparaît dans le segment de la LS. Le processus de constitution de sous-corpus exploite le résultat de l'alignement phrastique effectué sur le corpus EUROPARL. Les unités de traduction peuvent contenir de 1 à 2 phrases par langue. En raison du nombre variable de phrases par langue, nous parlerons de segments source et de segments cible mis en relation au sein d'une unité de traduction, comme dans le cas de l'alignement phrastique du corpus d'apprentissage. Le corpus étant aligné au niveau des phrases, les unités de traduction constituées par des phrases en anglais contenant de nouvelles instances des mots polysémiques et leurs traductions en grec, ont été automatiquement extraites. Des sous-corpus ont été ainsi construits, tant pour les 10 mots polysémiques pour lesquels nous avons procédé à un repérage manuel des équivalents que pour les 150 mots retenus du lexique bilingue anglais-grec, issu des résultats de l'alignement lexical automatique du corpus d'apprentissage. Ces sous-corpus ont la même forme que les sous-corpus construits à partir du corpus d'apprentissage (cf. figure 7, §1.3.10.). Les sous-corpus constitués pour un mot polysémique seront dorénavant appelés **sous-corpus de test** du mot polysémique.

2.2.5. Filtrage des sous-corpus en fonction des équivalents

Le sous-corpus correspondant à un mot polysémique comprend les unités de traduction où ce mot apparaît au sein du segment source. Les différentes instances du mot (*m*) au sein du sous-corpus sont traduites par des équivalents de traduction différents dans les segments de la LC. Le sous-corpus constitué

pour un mot polysémique a été filtré en fonction de ses équivalents. Ce filtrage est effectué de manière identique à celui du corpus d'apprentissage (§1.3.11.). La liste des équivalents possibles de chaque mot est extraite du lexique bilingue constitué lors de l'apprentissage et les unités de traduction sont triées en fonction de la présence d'un de ces équivalents en elles (cf. figure 8, §1.3.11.). Bien évidemment, l'utilisation des ensembles d'équivalents construits à partir du corpus d'apprentissage ne signifie pas que les mots étudiés ne peuvent pas être traduits par d'autres mots de la LC au sein du corpus de test. Cependant, ce qui nous intéresse est d'évaluer la performance de la méthode de désambiguïsation et de prédiction de traduction qui utilise les informations trouvées dans le corpus d'apprentissage. Les instances des mots qui sont traduits par d'autres équivalents dans le corpus de test ne sont, par conséquent, pas prises en compte.

Lors du filtrage effectué, les unités de traduction dont le segment de la LC contient plus d'un des équivalents possibles du mot polysémique sont éliminées, de la même manière que lors du filtrage des sous-corpus générés à partir du corpus d'apprentissage. La raison en est qu'un des équivalents peut traduire un autre mot de la phrase source, sans que nous puissions connaître celui qui traduit effectivement l'instance du mot source étudié. L'équivalent traduisant l'instance du mot polysémique dans les unités de traduction retenues constitue la **traduction de référence**. Celle-ci sera comparée à la proposition faite par notre méthode de sélection lexicale, ce qui permettra d'en évaluer la performance.

CONCLUSION

Dans ce chapitre, nous avons présenté les étapes de prétraitement des corpus d'apprentissage et d'évaluation qui seront utilisés dans la suite de ce travail. Cette chaîne de prétraitement peut être complètement automatisée, si la qualité des lexiques bilingues générés à l'issue de l'alignement des mots est bonne, tant en matière de précision que de rappel. Mais, ne disposant pas de résultats de très bonne qualité concernant ces deux aspects, nous avons introduit une étape de repérage manuel de traductions, dont les données nous serviront pour présenter en détail le fonctionnement des méthodes que nous proposons.

Une autre intervention manuelle a eu lieu au niveau du filtrage des entrées des lexiques bilingues automatiquement générés, dans le but d'en retenir un sous-ensemble (150 entrées) contenant des informations de traduction de bonne qualité. Ces étapes manuelles auraient pourtant pu être évitées si la quantité et la qualité des informations incluses dans les lexiques automatiquement générés étaient plus importantes.

Dans le chapitre suivant, nous allons présenter la manière dont les informations retenues du corpus d'apprentissage au bout de cette chaîne de traitement sont exploitées par notre méthode d'acquisition de sens. Néanmoins, avant d'en arriver à la présentation de cette méthode, nous allons décrire les résultats obtenus par l'application, au sein du cadre traductionnel dans lequel nous nous situons, d'une méthode d'acquisition de sens développée pour le traitement dans un cadre monolingue. Nous analyserons les problèmes que nous avons rencontrés lors de cette première expérience au niveau de la modélisation des correspondances inter-langues et nous expliquerons pourquoi les distinctions sémantiques obtenues par cette méthode ne seraient pas conformes à la désambiguïsation et à la prédiction de traduction dans un cadre bilingue. Les questions que nous nous sommes posées au moment de cette première expérience nous ont aidé à mieux comprendre les problèmes et à procéder, ainsi, au développement de la méthode que nous proposons finalement. Cette méthode d'acquisition de sens orientée vers la traduction sera présentée au chapitre suivant.

CHAPITRE 5

ACQUISITION DE SENS DANS UN CADRE MONOLINGUE

INTRODUCTION

Malgré le statut de la désambiguïsation lexicale (WSD) en tant que tâche intermédiaire, la majeure partie des méthodes de WSD développées jusqu'à présent sont indépendantes d'une application particulière, et il en va de même pour les inventaires sémantiques utilisés. Dans ce chapitre, nous explorerons l'applicabilité, dans un cadre de traduction, d'une méthode d'acquisition de sens développée dans un cadre monolingue, ainsi que la possibilité d'utilisation de ses résultats pour la WSD dans un cadre de traduction.

Le rôle de la WSD au sein d'une application traductionnelle consiste à **identifier le sens** véhiculé par de nouvelles instances des mots ambigus[248] de la

[248] Le mot ambigu à analyser est souvent appelé **mot cible** dans les travaux d'acquisition de sens et de désambiguïsation. Nous avons choisi de ne pas utiliser ce terme dans le cadre bilingue dans

LS – dans les cas où cela serait utile – dans un but de **prédiction de traductions** sémantiquement pertinentes. Cette tâche peut impliquer la création de correspondances entre mots source et équivalents (EQVs) au niveau du sens, ce qui implique, à son tour, le repérage des éléments des deux langues qui seraient mis en relation. Ces éléments seraient les sens des mots ambigus, du côté de la LS, et du côté de la LC, leurs EQVs de traduction. Par conséquent, la création de correspondances sémantiques présuppose, d'une part, le repérage des sens véhiculés par un mot ambigu et, d'autre part, le repérage de ses EQVs possibles.

Dans la suite de ce chapitre, nous décrirons l'expérience que nous avons menée à propos de l'exploitation, pour la WSD et la prédiction de traductions, des correspondances inter-langues créées sur la base des sens repérés par une méthode monolingue. Nous étudierons le fonctionnement de cette méthode et nous analyserons la nature des correspondances établies pour l'échantillon de mots étudié. Nous présenterons ensuite les résultats du processus de WSD et de prédiction de traductions, ainsi que les conclusions tirées de cette expérience concernant l'applicabilité d'une telle méthode monolingue dans un cadre de traduction.

1. Acquisition de sens au niveau de la langue source

1.1. Présentation de la méthode d'acquisition de sens

1.1.1. Principes sous-jacents

L'expérience décrite ici concerne l'application d'une **méthode contextuelle d'acquisition de sens** utilisant des informations monolingues à un échantillon de données de notre corpus parallèle. Plus précisément, il s'agit de repérer les sens des mots ambigus source par application de la méthode sur des données de la LS de notre corpus d'apprentissage, sans tenir compte des informations de traduction. La méthode utilisée se situe dans la même ligne que d'autres

lequel ce travail se situe, afin d'éviter une confusion avec l'utilisation du terme « cible » pour des informations provenant du côté de la langue cible, c'est-à-dire des traductions.

méthodes contextuelles de repérage de sens et de désambiguïsation développées dans un contexte monolingue (Schütze, 1998 ; Véronis, 2003, 2004 ; Agirre et Soroa, 2007b ; Klapaftis et Manandhar, 2007). L'approche adoptée dans ces méthodes contextuelles relève de l'**approche contextuelle du sens** (Wittgenstein, 1953 : 18 ; Harris, 1954 ; Firth, 1957a,b), selon laquelle le **sens** des mots correspond à leur **usage** dans les textes. En ce qui concerne les mots ambigus, il est considéré que leurs sens peuvent être distingués en regroupant leurs usages similaires, regroupement qui peut s'effectuer sur la base de la similarité de leurs contextes lexicaux.

Plus précisément, la méthode utilisée ici s'inspire, et est par conséquent très proche, de celle proposée par Véronis (2003, 2004). Le repérage de sens au sein de cette méthode se fait à l'aide de **graphes de cooccurrence** élaborés à partir des contextes lexicaux des mots. Ces graphes décrivent les différents sens des mots ambigus à l'aide de leurs cooccurrents au sein des textes du corpus. Dans la méthode initiale (Véronis, *ibid*.), un graphe est construit pour un mot ambigu. L'algorithme **HyperLex** permet de détecter des **composantes de haute densité** au sein du graphe correspondant à un mot, composantes qui sont interprétées comme les différents sens du mot. L'étude que nous menons ici se focalise sur l'analyse d'un petit ensemble de mots polysémiques de la LS[249]. L'objectif est de créer des graphes correspondant à ces mots précis et non de donner une image de la sémantique de l'ensemble des mots du corpus. Le but visé est d'estimer la possibilité de création de correspondances entre les sens des mots ambigus source, repérés par cette méthode, et leurs EQVs de traduction dans la LC, utilisables dans un but de prédiction de traductions.

[249] L'application envisagée pour la méthode d'acquisition de sens et de désambiguïsation proposée par Véronis (2003, 2004) est la Recherche d'Information sur le Web. Néanmoins, dans la présentation de sa méthode, l'analyse se focalise également sur un petit ensemble de mots polysémiques.

1.1.2. Données utilisées

La méthode utilisée est dirigée par les données ; les caractéristiques du corpus utilisé pour l'entraînement sont décrites dans le chapitre 4.

Les mots que nous avons analysés sont des noms polysémiques de la LS ayant un grand nombre d'EQVs de traduction possibles. Les EQVs sont trouvés dans les entrées correspondant aux mots source, au sein du lexique bilingue anglais-grec manuellement construit. Les noms qui ont été étudiés et leurs EQVs sont décrits dans le Tableau 1.

mot ambigu	EQVs de traduction
movement	*κυκλοφορία* (kykloforia)[250], *διακίνηση* (diakinisi), *κίνηση* (kinisi), *μετακίνηση* (metakinisi), *κινητικότητα* (kinitikotita),*κίνημα* (kinima)
class	*τάξη* (taksi), *κατηγορία* (katigoria), *μάθημα* (mathima), *εκπαίδευση* (ekpaidefsi), *τμήμα* (tmima), *σώμα* (soma)
competence	*αρμοδιότητα* (armodiotita), *ικανότητα* (ikanotita), *επάρκεια* (eparkeia), *δικαιοδοσία* (dikaiodosia), *δεξιότητα* (deksiotita), *δυνατότητα* (dynatotita), κύρος (kyros)
capital	*κεφάλαιο* (kefalaio), *πρωτεύουσα* (protevousa), *δυναμικό* (dynamiko), *κιονόκρανο* (kionokrano), *χώρα* (chora)
passage	*διέλευση* (dielefsi), *καλλιέργεια* (kaliergeia), *πέρασμα* (perasma), *πρόσβαση* (prosvasi), *χωρίο* (chorio)
plant	*φυτό* (fyto), *εγκατάσταση* (egkatastasi), *μονάδα* (monada), *σταθμός* (stathmos), *εργοστάσιο* (ergostasio)
occupation	*απασχόληση* (apascholisi), *ασχολία* (ascholia), *δραστηριότητα* (drastiriotita), *επάγγελμα* (epangelma), *κατάληψη* (katalipsi), *κατοχή* (katochi)

Tableau 1. Mots ambigus étudiés et leurs EQVs de traduction

Le nombre de mots analysés n'est pas important car, dès les premiers résultats obtenus, nous avons repéré les faiblesses de la méthode et nous l'avons donc abandonnée pour une autre, plus conforme au traitement dans un cadre bilingue[251]. Nous illustrerons le fonctionnement et les résultats de la méthode à l'aide de trois exemples, qui concernent les mots : *movement*, *class* et *competence*.

[250] La translittération des caractères grecs en caractères latins, au sein de cette thèse, se fonde sur le standard ISO 843:1997 TR. Le tableau de translittération est donné en Annexe C.1.
[251] Cette méthode sera présentée dans le chapitre suivant.

1.1.3. Informations retenues à partir du corpus d'apprentissage

La méthode contextuelle utilisée pour le repérage des sens d'un mot source fait appel, comme nous l'avons déjà dit, à des informations de **cooccurrence lexicale**. Le corpus d'apprentissage étant divisé en **unités de traduction** qui contiennent des segments des deux langues mis en correspondance lors de l'alignement des phrases (cf. §1.2.2, chapitre 4), le contexte d'une instance du mot source est délimité par le **segment de la LS** qui la contient. Le sous-corpus correspondant à un mot ambigu est construit de la manière décrite dans le paragraphe 2.3.10 du chapitre 4.

Les informations de cooccurrence retenues pour un mot ambigu concernent, plus précisément, les **noms** et les **adjectifs** qui apparaissent avec lui au sein des segments de la LS du sous-corpus, ramenés à leurs **lemmes**. Dans le cas de l'unité de traduction décrite dans la figure 1, par exemple, les informations contextuelles retenues, pour cette instance du mot *movement*, concernent les formes des noms et des adjectifs[252] qui apparaissent au sein du segment anglais ('seg.AN').

<seg id="seg.AN…"> The effect of the application of this paragraph shall not result in conditions for the temporary **movement** of workers in the context of the transnational provision of services between Germany or Austria and Lithuania which are more restrictive than those prevailing on the date of signature of the Treaty of Accession. **</seg>**

<seg id="seg.GR..."> Η εφαρμογή της παρούσας παραγράφου δεν μπορεί να δημιουργήσει συνθήκες προσωρινής κυκλοφορίας των εργαζομένων στο πλαίσιο της διεθνικής παροχής υπηρεσιών μεταξύ της Γερμανίας ή της Αυστρίας και της Λιθουανίας, οι οποίες είναι πιο περιοριστικές από αυτές που επικρατούν την ημερομηνία υπογραφής της Συνθήκης Προσχώρησης. **</seg>**

Figure 1. Unité de traduction regroupant deux segments

Les informations contextuelles de cette unité, qui seront exploitées par la méthode d'acquisition de sens, sont décrites par l'ensemble de mots suivant : {*effect, application, paragraph, condition, temporary, worker, context, transnational, provision, service, Germany, Austria, Lithuania, restrictive, date, signature, Treaty, Accession*}.

[252] Les verbes ne sont pas pris en considération lors de cette expérience, en raison de l'influence négative sur les résultats de leur fort degré de polysémie (Véronis, 2004).

1.1.4. Objectifs de l'analyse

Les buts principaux de l'analyse effectuée pour chaque mot ambigu de notre échantillon sont les suivants :

a. analyser la **sémantique** du **mot source**, c'est-à-dire repérer les sens véhiculés par ce mot au sein du corpus d'apprentissage

b. repérer des **liens** entre les **sens** repérés et les **équivalents** de traduction du mot.

Le premier point permet de repérer les éléments de la LS qui participeront à la création des correspondances sémantiques inter-langues. Le deuxième point permet précisément la mise en place de ces correspondances, en créant des liens entre les sens des mots ambigus source et leurs équivalents de traduction dans la LC. Dans le paragraphe suivant, nous allons décrire le processus de repérage des sens des mots ambigus source.

1.2. Construction de graphes de cooccurrence décrivant les sens du mot ambigu

1.2.1. Nature des graphes construits

Les sens d'un mot ambigu sont mis en évidence par la construction des **graphes de cooccurrence** qui lui correspondent. L'algorithme utilisé permet la création d'**un graphe** pour **chacun des sens** du mot étudié. Par conséquent, l'ensemble des graphes construits pour un mot est supposé décrire sa sémantique. Ceci constitue un point de divergence entre l'algorithme utilisé ici et l'algorithme HyperLex (Véronis, 2003, 2004). HyperLex repère des composantes de haute densité au sein d'**un seul graphe**, qui correspond au mot ambigu ; les composantes ainsi repérées correspondent aux différents sens du mot. La divergence entre les deux algorithmes est due au fait que HyperLex construit le graphe en utilisant la **matrice de cooccurrence** créée pour le **mot ambigu**, à partir de l'ensemble des contextes qui lui correspondent (les segments LS de son sous-

corpus). En revanche, nous utilisons ici la **liste de fréquence** construite pour **le mot ambigu** et la **matrice de cooccurrence** construite pour **le mot le plus fréquent** de la liste de fréquence du mot ambigu.

Le mot le plus fréquent diffère à chaque itération de l'algorithme utilisé, comme nous le montrerons plus loin, en raison de l'élimination d'un ensemble de mots de cette liste à chaque itération. Les **nœuds** des graphes construits ici pour chacun des sens du mot ambigu correspondent aux mots qui co-apparaissent avec lui au sein des contextes retenus. Des **arêtes** lient deux mots (nœuds) lorsqu'ils apparaissent ensemble, de manière significative, à proximité du mot ambigu.

1.2.2. Données utilisées

Le sous-corpus constitué pour le mot ambigu, qui contient les segments où il apparaît et leurs traductions, est divisé en deux parties, dont l'une est réservée à l'apprentissage et l'autre à l'évaluation[253]. La partie de l'évaluation, qui constitue 20 % du sous-corpus, est mise à part dès le début. Une liste de fréquence est construite à partir de la partie du sous-corpus réservée à l'apprentissage, qui contient les fréquences d'occurrence des mots qui apparaissent dans le contexte du mot ambigu. Les calculs portent sur les **formes** des mots du contexte (*types*) et non sur les occurrences (*tokens*). Les textes du corpus d'apprentissage étant lemmatisés, les mots trouvés dans les contextes peuvent être facilement ramenés à leur lemme. Ainsi, une liste de fréquence ne contient pas les fréquences d'occurrence des mots mais les **fréquences cumulées** associées aux lemmes. La fréquence cumulée d'un lemme correspond à la somme des fréquences des occurrences (ou formes fléchies) associées à ce lemme et trouvées dans les contextes[254]. L'utilisation des lemmes aide à diminuer l'effet de la **dispersion des données**, phénomène ayant une influence négative sur les résultats des méthodes distributionnelles.

[253] Cette division est opérée seulement dans le cadre de cette première expérience. Dans les expériences qui concernent la méthode d'acquisition de sens et de désambiguïsation qui sera proposée par la suite, le corpus INTERA est utilisé pour l'apprentissage et le corpus EUROPARL pour l'évaluation.

[254] Par exemple, si les occurrences *teacher* et *teachers* sont trouvées *5* et *6* fois respectivement dans les contextes du mot ambigu, la fréquence du lemme *teacher* dans la liste sera de *11*.

1.2.3. Description du processus de construction des graphes

De la liste de fréquence construite pour un mot ambigu, seuls les mots ayant une fréquence supérieure à 2 sont retenus. L'algorithme de construction des graphes sélectionne, tout d'abord, le mot le plus fréquent dans la liste. Une **matrice de cooccurrence** (M) est construite à partir de l'ensemble des contextes où ce mot apparaît au sein du sous-corpus du mot ambigu. Les éléments de la matrice sont les lemmes auxquels sont ramenés les noms et les adjectifs des contextes du mot le plus fréquent. Chaque case M[*i*][*j*] de la matrice de cooccurrence contient la fréquence de cooccurrence du lemme qui se trouve en position *i* et du lemme qui se trouve en position *j*, dans l'ensemble des contextes. Un exemple de la matrice construite pour le mot *free* (le mot le plus fréquent dans la liste de fréquence de *movement*) est donné dans le Tableau 2 : la fréquence de cooccurrence des mots *market* et *transnational*, dans les contextes de *free* (au sein du sous-corpus de *movement*), est de 8 ; celle de *market* et de *right* est de 10 ; celle de *region* et de *national* est de 16 ; etc. Seuls les cooccurrents ayant une fréquence supérieure à 1 sont retenus à partir de cette matrice[255].

	transnational	**market**	**region**	**right**	**national**
transnational	-	8	8	7	15
market	8	-	8	10	17
region	8	8	-	7	16
right	7	10	7	-	16
national	15	17	16	16	-

Tableau 2. Echantillon de la matrice de cooccurrence de *free*

Le graphe correspondant au cooccurrent le plus fréquent du mot source est ensuite construit à l'aide de sa matrice de cooccurrence et de la liste de fréquence du mot ambigu. Les nœuds de ce graphe sont les mots retenus après le filtrage des contextes, en fonction des parties du discours et des fréquences d'occurrence

[255] Les seuils de fréquence et de cooccurrence (2 et 1 respectivement) sont bas en raison de la petite taille du corpus d'apprentissage.

et de cooccurrence. Les arcs liant les nœuds sont pondérés par la formule suivante :

$$w_{A,B} = 1 - \max[p(A|B), p(B|A)] \qquad (1.2.3.1)$$

où $p(A|B)$ est la **probabilité conditionnelle** d'observer l'élément A dans un contexte où l'élément B apparaît et, inversement, $p(B|A)$ est la probabilité conditionnelle d'observer l'élément B dans un contexte où A apparaît.

Les probabilités sont estimées à l'aide des fréquences d'occurrence et de cooccurrence des mots :

$$p(A|B) = f_{A.B}/f_B \quad (1.2.3.2.) \qquad \text{et} \qquad p(B|A) = f_{A.B}/f_A \quad (1.2.3.3.)$$

Par exemple, les probabilités conditionnelles des mots *transnational* et *sector*, qui ont respectivement une fréquence de 17 et de 9 dans le sous-corpus du mot ambigu *movement*, et une fréquence de cooccurrence de 8 dans les segments correspondant au mot le plus fréquent dans la liste de fréquence de *movement* (*free*) sont les suivantes :

p(*transnational* | *sector*) = 8/9 = **0,888** et p(*sector* | *transnational*) = 8/17 = **0,47**

Le poids $w_{A,B}$ de ces deux mots est :

$$w_{transnational,sector} = 1 - \max [p(transnational \mid sector), p(sector \mid transnational)]$$

$$w_{transnational,sector} = 1 - 0{,}888 = \mathbf{0{,}112}$$

Le poids $w_{transnational,sector}$ est attribué à l'arc qui lie les nœuds correspondant à *transnational* et à *sector*, au sein du graphe de *free*. Ce poids reflète la **distance sémantique** entre les mots : quand il vaut 0, les mots sont toujours associés tandis que lorsqu'il vaut 1, ils ne le sont jamais. Les arcs retenus ont un poids <0,9 et sont considérés comme liant les nœuds qui se trouvent en relation de **cooccurrence significative**. Une fois le premier graphe construit, le mot le plus fréquent et tous ses voisins dans le graphe sont éliminés de la liste de fréquence

du mot ambigu. Le mot le plus fréquent de la liste résultante est sélectionné et la procédure continue de la même manière : les contextes correspondant à ce mot sont retenus, la matrice de cooccurrence est construite, le graphe correspondant est créé et le mot en question ainsi que ses voisins sont, par la suite, éliminés de la liste de fréquence. Ce processus est réitéré tant que le mot le plus fréquent dans la liste de fréquence a au moins 6 voisins propres[256]. L'algorithme utilisé est décrit de manière plus formelle (en pseudo-code) en Annexe C.

1.2.4. Analyse des résultats

L'hypothèse sous-jacente à ce processus est que les différents sens du mot ambigu sont décrits par les **petits graphes** qui lui correspondent, construits un par un. Cette hypothèse est similaire à celle qui sous-tend la méthode utilisée par Véronis (*ibid*.), où le repérage des différents sens d'un mot se fait par la détection de composantes de haute densité dans le graphe du mot. La différence principale est, comme nous l'avons déjà dit, que dans la méthode de Véronis, les composantes sont repérées au sein du **graphe global** construit pour le mot.

Mots ambigus	Equivalents[257]	Corpus d'entraînement	Sens	Nombre de nœuds	Nombre d'arêtes	Densité du graphe
movement	κυκλοφορία(251), διακίνηση(38), κίνηση(28), μετακίνηση(19), κίνημα(11), κινητικότητα(6)	353	free	245	2177	0,07
			freedom	95	447	0,1
			relation	23	113	0,44
			restriction	14	23	0,25
class	τάξη(201), κατηγορία(20), μάθημα(9), τμήμα(3), εκπαίδευση(2), σώμα(1)	236	school	125	467	0,06
			number	23	30	0,11
			device	17	40	0,29
			lesson	11	21	0,38
competence	αρμοδιότητα(118), ικανότητα(88), επάρκεια(4), δικαιοδοσία(4),δεξιότητα(3) δυνατότητα(2), κύρος(1)	220	member	143	617	0,06
			skill	29	54	0,13
			qualification	12	19	0,28

Tableau 3. Informations quantitatives sur les mots étudiés

[256] Les seuils utilisés sont ceux proposés par Véronis (*ibid*.). Nous avons expérimenté d'autres seuils, sans remarquer d'amélioration des résultats.

[257] Ordonnés en fonction de leur fréquence dans le corpus d'entraînement.

Le Tableau 3 contient des informations quantitatives sur les mots ambigus étudiés. La cinquième colonne du tableau décrit les sens repérés pour chaque mot ambigu. Chaque sens est désigné par le mot qui déclenche la construction du graphe correspondant. Ce mot est, comme nous l'avons déjà montré, le mot le plus fréquent dans la liste de fréquence du mot ambigu, à chaque itération de l'algorithme, qui satisfait un certain nombre de conditions. Ce mot correspond, par ailleurs, au nœud ayant le plus haut **degré** dans le graphe construit. Dans le cas du mot *movement*, par exemple, l'algorithme repère quatre sens décrits par les mots suivants : *free*, *freedom*, *relation*, *restriction*. La troisième colonne du tableau décrit la taille du corpus d'entraînement pour le mot source, en nombre de segments. Les nombres de nœuds et d'arcs du graphe décrivant chaque sens sont également décrits dans le tableau (colonnes 6 et 7). La **densité** du graphe correspond au rapport des arêtes existantes au nombre maximal d'arêtes possibles, étant donné le nombre de nœuds.

1.3. Etude de la nature des sens proposés

1.3.1. Granularité et nombre des sens

Une remarque générale s'impose à propos des résultats de la méthode de repérage de sens : les **sens** proposés sont de **granularité trop fine** et, dans certains cas, pourraient même être plus convenablement caractérisés comme **usages**[258]. Ceci est évident, par exemple, dans le cas de *movement* : chacun des quatre sens repérés pour ce mot est décrit dans le Tableau 4, à l'aide du mot qui déclenche la construction du graphe correspondant et des voisins qui lui sont le plus fortement liés.

[258] Ceci est d'ailleurs signalé par Véronis (*ibid*.), qui préfère de parler d' « usages » des mots que de « sens ».

mot ambigu	sens	voisins
movement	free	student, goods, worker, barrier, citizen, capital, …
	freedom	border, residence, territory, immigration…
	relation	Sweden, Finland, Belgium, Italy, Spain…
	restriction	animal, disease, trade, health, risk...

Tableau 4. Sens repérés pour le mot *movement*

Les deux premiers sens obtenus pour *movement*, décrits respectivement par les mots *free* et *freedom*, sont relatifs à la liberté de mouvement ou de circulation (de personnes, d'articles, de capitaux, etc.) au sein de l'Union Européenne. En observant les segments où *movement* co-apparaît avec ces deux mots, nous ne constatons pas de différence de sens considérable. Le troisième sens est décrit par le mot *relation*, qui appartient pourtant toujours à l'expression figée *in relation to*, au sein du sous-corpus de *movement*. Le problème, dans ce cas, est évidemment que le mot est considéré séparément et non en tant que partie de cette expression ; par conséquent, le troisième sens ne constitue pas non plus un sens pertinent. Enfin, le sens décrit par le mot *restriction* fait référence à la circulation de maladies (par ex. transmises par les animaux) entre les pays de l'Union Européenne. Nous estimons qu'il est préférable de considérer cette distinction comme un usage différent du sens de « circulation », exprimé par les deux premiers sens, qui pourraient être également considérés comme des usages différents du même sens.

Nous remarquons donc, d'une part, la forte similarité des trois sens proposés et, d'autre part, le non repérage d'un autre sens effectivement distinct, véhiculé par *movement* dans le corpus, à savoir le sens de « mouvement social ». Ce manque est dû au petit nombre d'instances du mot véhiculant ce sens, qui rend la quantité des informations contextuelles permettant son repérage trop petite. L'algorithme ne parvient pas à détecter le sens en question parce qu'aucun des mots du contexte qui permettraient sa description ne remplit les conditions nécessaires à la construction d'un graphe distinct (à savoir, une fréquence et un nombre de voisins suffisants).

Une autre remarque qui s'impose concerne la facilité selon laquelle le nombre de sens proposés, leur granularité ainsi que les mots qui servent à les décrire peuvent changer, en fonction des seuils utilisés. Nous avons en effet constaté que les résultats obtenus différent largement, même avec une modification minime des seuils, changement qui ne s'accompagne pourtant pas d'une amélioration qualitative. Les résultats décrits ici ont été obtenus en utilisant les seuils qui donnaient les meilleurs résultats pour un ensemble de mots polysémiques (seuils décrits dans le paragraphe précédent).

La **granularité trop fine** des sens proposés et la proposition de **sens non pertinents**, qui reflètent seulement des usages différents des mots, constituent des inconvénients connus des **méthodes de cooccurrences**, qui font appel uniquement à des **informations de surface**, contenues dans les textes, et non à des informations plus abstraites[259]. Ce problème est fortement lié au phénomène de la **dispersion des données**, fort présent dans le cadre de méthodes basées sur les cooccurrences de premier ordre (cf. §1.2.3, chapitre 2). Ceci est dû au fait que des instances sémantiquement similaires des mots apparaissent dans des contextes où des mots conceptuellement similaires (mais non lexicalement) sont utilisés (Purandare et Pedersen, 2004a,b)[260]. Lors de l'application de la méthode, nous avons remarqué que la dissimilarité des contextes où le mot véhicule le même sens provoque la constitution de graphes décrivant des usages différents de ce sens, qui ne peuvent être regroupés en raison précisément de la dissimilarité des informations de cooccurrence.

1.3.2. Regroupement des sens obtenus

Etant donné la finesse de granularité des distinctions sémantiques obtenues, qui correspondent à des **usages**, leur regroupement peut constituer une solution pour aboutir à des **distinctions plus grossières**, correspondant à des **sens**. Dans un cadre bilingue, nous avons envisagé la possibilité de parvenir à ce but en

[259] Ces informations pourraient être, par exemple, des informations sur les classes des mots, technique utilisée dans le cadre de la désambiguïsation lexicale supervisée pour diminuer l'effet de la dispersion des données, comme nous l'avons déjà souligné dans le paragraphe 3.2, chapitre 3.

[260] Pour cette raison, Purandare et Pedersen (2004a,b) proposent d'utiliser les cooccurrences de deuxième ordre.

établissant des **correspondances sémantiques** inter-langues. Ces correspondances seraient plus correctement caractérisées comme des correspondances **usages-équivalents** (et non **sens-équivalents**). Dans le paragraphe suivant, nous allons décrire une manière d'établir de telles correspondances et examinerons, par la suite, la possibilité d'utiliser ces correspondances afin de diminuer la granularité des distinctions sémantiques obtenues.

2. Création de correspondances inter-langues

2.1. Traits caractérisant les sens-usages obtenus

Les « sens-usages » des mots ambigus sont révélés, comme nous l'avons montré, à l'aide d'informations provenant des contextes lexicaux de leurs instances. Plus précisément, les graphes décrivant les différents sens-usages des mots sont construits sur la base des informations de cooccurrence lexicale relatives à leurs instances. Ces informations se retrouvent dans les graphes construits et peuvent servir à mettre les sens-usages en correspondance avec les EQVs de traduction des mots ambigus.

La modélisation de correspondances inter-langues, à l'aide d'informations contextuelles provenant de la LS, pourrait être considérée comme une définition de critères conditionnant l'utilisation des EQVs possibles du mot source en tant que traductions de ses instances en contexte[261]. Ainsi, les correspondances construites sur la base d'informations contextuelles relatives au mot ambigu permettent, d'une part, la **désambiguïsation** d'une nouvelle instance du mot par prise en compte de son contexte et, d'autre part, la **sélection de l'EQV le plus adéquat** pour cette instance.

[261] Les correspondances construites de cette manière pourraient être conçues comme décrivant les facteurs contextuels qui déterminent les probabilités conditionnelles des équivalences (Catford, 1965 : 29). Cette possibilité de décrire la variabilité présente dans la traduction en des termes probabilistes et celle de déterminer la probabilité statistique des relations d'équivalence par référence au contexte, constituent des avantages certains de l'utilisation de corpus parallèles dans une étude traductionnelle (Malmkjær, 1998).

2.2. Traits caractérisant les équivalents

L'étape de mise en relation des sens-usages et des équivalents présuppose le repérage des EQVs possibles du mot ambigu au sein du corpus d'apprentissage. Les correspondances sont établies à l'aide des contextes de la LS correspondant à chacun de ces EQVs. Le **contexte de la LS** correspondant à un **EQV** est composé des mots qui apparaissent autour du mot ambigu (dans les segments de la LS) dans l'ensemble des unités de traduction où il est traduit par cet EQV précis, au sein du corpus d'apprentissage. De ces contextes, nous ne retenons que les noms et les adjectifs. Les groupes de segments de la LS correspondant à chaque EQV sont formés à partir du sous-corpus du mot ambigu, de la manière décrite dans le paragraphe 2.3.11, chapitre 4.

2.3. Graphes de cooccurrence des équivalents

Les informations contextuelles de la LS relatives à chacun des EQVs constituent l'entrée d'un calcul similaire à celui employé pour la construction des graphes du mot source. Une matrice de cooccurrence est construite à partir des segments correspondant à chaque équivalent. Le poids $w_{A,B}$ entre chaque paire de cooccurrents (donné par la formule 1.2.3.1.) est calculé en utilisant la **matrice** correspondant à l'**EQV** et la **liste de fréquence** du **mot ambigu** (il s'agit de la liste de fréquence initiale, construite avant la création des graphes de cooccurrence au niveau de la LS). Ainsi, les mots de la LS qui ont une relation de cooccurrence significative avec l'EQV sont retenus. De cette manière, un **graphe de cooccurrence** est construit pour **chacun des EQVs** de traduction comprenant ses cooccurrents dans la LS, c'est-à-dire les mots qui apparaissent à proximité du mot ambigu de manière significative lorsqu'il est traduit par cet EQV précis.

2.4. Estimation de similarité des graphes source et cible

2.4.1. Calcul de recouvrement des graphes

Des correspondances inter-langues peuvent être établies pour le mot ambigu au niveau des graphes décrivant les sens-usages du mot et les graphes de ses EQVs. Les correspondances entre les graphes sont établies sur la base de leur similarité, qui est estimée en termes de **partage de traits** (Tversky, 1977). Plus précisément, la similarité est obtenue en calculant le taux de **recouvrement** entre le graphe de chaque EQV et les graphes décrivant les différents « sens-usages » du mot ambigu. Les traits sur lesquels porte ce calcul ne sont pas les cooccurrents individuels – qui correspondent aux nœuds du graphe – mais les **paires de nœuds** liés, qui représentent des relations significatives entre les cooccurrents. Si, par exemple, les mots *a* et *b* sont liés dans le graphe d'un EQV, des nœuds correspondant à *a* et à *b* sont recherchés dans le graphe du mot source et, si ces nœuds existent, il reste à vérifier si une arête lie les deux. Si tel n'est pas le cas, les mots ne sont alors pas retenus.

Nous avons en effet constaté que le calcul de recouvrement entre cooccurrents individuels introduit de faux liens entre les graphes des EQVs et les graphes source, en raison de l'ambiguïté des cooccurrents. Par contre, la prise en compte des relations que ces mots entretiennent avec d'autres cooccurrents restreint leur ambiguïté. Les liens retenus de cette manière s'avèrent donc nettement plus pertinents. Le pseudo-code de l'algorithme utilisé pour calculer le recouvrement entre le graphe d'un EQV et chaque graphe de la LS est décrit en Annexe C.

La correspondance entre un sens-usage et un EQV est donc décrite par l'ensemble des **associations de traits contextuels** qui leur sont communes. Un sous-ensemble des associations communes à la première composante de *movement* (décrite par le mot *free*) et au graphe de l'EQV *κυκλοφορία* (kykloforia) est décrit dans la figure suivante. Le score attribué aux associations correspond au poids $w_{A,B}$, calculé lors de la construction du graphe de l'EQV.

agreement--order = 0.62	agreement--worker = 0.19
free--individual = 0.21	citizenship--person = 0.43
market--sector = 0.12	free--national = 0.63
overall--strategy = 0	market--national = 0.24

…

Figure 2. Associations communes à *κυκλοφορία* et au premier sens-usage de *movement*

2.4.2. Description des résultats obtenus

Le Tableau 5 décrit les correspondances établies entre les sens-usages d'une partie des mots ambigus étudiés et leurs EQVs en grec. Dans la deuxième colonne, les sens-usages d'un mot ambigu sont désignés par le mot qui déclenche la construction du graphe correspondant. La troisième colonne contient certains voisins directs de ce mot dans le graphe, qui servent à illustrer les différents usages. Par exemple, les voisins du mot *device*, qui décrit un usage de *class*, montrent qu'il est question dans les textes de « catégories d'appareils médicaux ». La dernière colonne du tableau contient les EQVs grecs correspondant à chaque sens-usage, comme cela a été démontré par le calcul de recouvrement.

Notons que plusieurs EQVs peuvent être mis en relation avec un seul sens-usage et, à l'inverse, un EQV peut être lié à plusieurs sens-usages. Dans le premier cas, le calcul de recouvrement établit des correspondances entre des sous-ensembles d'un graphe de la LS et des graphes correspondant à des EQVs différents tandis que, dans le deuxième cas, des correspondances sont établies entre des sous-ensembles du graphe d'un EQV et des graphes de sens différents.

Mot ambigu	Sens-usages	Voisins	Equivalents
class	school	classroom, teacher, pupil, elementary…	τάξη, μάθημα, εκπαίδευση, τμήμα
	number	minimum, maximum, total, high, average…	τάξη, κατηγορία
	device	implant, breast, instruction, practice, medical…	κατηγορία
	lesson	orientation, written, second, basis…	τάξη
competence	member	state, sphere, infringement, power, exercise…	αρμοδιότητα, δικαιοδοσία, ικανότητα, κύρος, επάρκεια
	skill	personal, lifelong, language, mathematics…	ικανότητα, επάρκεια
	qualification	recognition, development, partner, trust…	αρμοδιότητα, κατάρτιση
movement	free	student, goods, worker, barrier, citizen…	κυκλοφορία, μετακίνηση, κίνηση, κινητικότητα, διακίνηση, κίνημα
	freedom	border, residence, territory, immigration…	μετακίνηση, διακίνηση, κίνηση, κινητικότητα, κυκλοφορία
	relation	Sweden, Finland, Belgium, Italy, Spain…	κίνηση, κυκλοφορία, κινητικότητα
	restriction	animal, disease, trade, health, risk...	κίνηση, διακίνηση

Tableau 5. Correspondances entre sens-usages et EQVs des mots ambigus

3. Désambiguïsation lexicale et prédiction de traductions

3.1. Exploitation des correspondances inter-langues

Les correspondances établies, lors de l'étape précédente, entre les sens-usages des mots ambigus et leurs EQVs peuvent être exploitées pour la désambiguïsation (WSD) de nouvelles instances des mots ambigus et pour la prédiction des traductions les plus adéquates de ces instances. Les informations utilisées pour ce processus sont constituées par les **ensembles de traits** (associations communes) qui permettent la mise en correspondance des graphes de la LS et de la LC. Ces traits, qui servent à décrire les correspondances en question, peuvent être comparés aux traits trouvés dans les nouveaux contextes. L'estimation de la similarité de ces ensembles de traits permet la sélection du sens le plus approprié du mot ambigu en contexte ainsi que la prédiction de sa traduction la plus appropriée. L'utilisation de ces correspondances permet en

effet de restreindre le choix entre les EQVs de traduction possibles d'un mot source. Les correspondances observées au niveau des mots, avant le repérage des sens, sont très grossières (*competence*, par exemple, a 7 équivalents de traduction possibles en grec), tandis que celles établies par le calcul de recouvrement entre sens-usages et EQVs sont plus fines (1 sens-usage correspond à 1-5 équivalents).

3.2. Traitement des nouveaux contextes

Dans le cas d'une nouvelle instance du mot ambigu, c'est le contexte lexical du mot qui nous guide dans le choix du sens et de l'EQV les plus corrects. Ce contexte est filtré afin de n'en préserver que les **noms** et les **adjectifs**. Puis, une liste contenant des **associations** entre ces éléments est construite, associations qui montrent leur **relation de cooccurrence** dans le segment de texte en question. Par exemple, dans le cas de la phrase d'entrée suivante :

> *The resource teacher prepares materials, which the class teacher can use if necessary*.

certaines associations retenues sont les suivantes : *resource-teacher*, *resource-material*, *resource-class*, *resource-necessary*, *teacher-resource*, *teacher-material*, *teacher-class*, *teacher-necessary*, etc. Cet ensemble d'associations décrit donc le contexte de la nouvelle instance du mot *class*.

3.3. Comparaison entre nouveaux contextes et correspondances inter-langues

L'ensemble des associations élaboré à partir d'un nouveau contexte est ensuite comparé aux résultats de l'étape précédente, c'est-à-dire aux ensembles de traits qui décrivent les correspondances entre les sens et les EQVs, constitués à partir du corpus d'entraînement. Des ensembles précédemment établis, nous ne retenons que celui(ceux) qui partage(nt) des traits avec l'ensemble des associations construit à partir du nouveau contexte. Il s'agit alors, encore une fois, d'un calcul de similarité en termes de **partage de traits**. Pour la phrase d'entrée citée dans le paragraphe précédent, l'ensemble retenu est celui qui décrit

la correspondance entre le sens-usage *school* du mot polysémique *class* et l'EQV *τάξη*. De cette manière, le sens selon lequel le mot ambigu source est utilisé dans ce nouveau contexte est sélectionné (*school*), ainsi que la traduction la plus adéquate à cette nouvelle occurrence (*τάξη*). La WSD et la prédiction de traduction sont donc faites simultanément.

Il se peut que des relations soient aussi repérées entre le contexte de la nouvelle instance et plusieurs des ensembles préétablis, mais ces relations ne sont pas très nombreuses (de 2 à 4). Ainsi, même si nous n'obtenons pas une proposition de traduction unique pour la nouvelle instance, nous parvenons néanmoins à restreindre les choix de traduction. Dans le cas de propositions multiples, il est également possible d'attribuer une **préférence** à un EQV et à un sens-usage, en fonction de la **quantité** et des **poids des associations communes** entre le nouveau contexte et les traits décrivant les correspondances inter-langues, sans exclure les autres EQVs.

Néanmoins, si le but consiste uniquement à prédire l'EQV de traduction correct – et non à désambiguïser la nouvelle instance du mot ambigu –, il est possible de n'utiliser que les graphes correspondant aux équivalents. Dans ce cas, l'**ensemble des associations** incluses dans le **graphe de chaque EQV** est comparé aux associations retenues à partir du nouveau contexte et celui(ceux) qui partage(nt) le plus de traits avec ce contexte est(sont) retenu(s).

4. Evaluation des processus de WSD et de prédiction de traduction

4.1. Données de test

Les deux processus de désambiguïsation et de prédiction de traduction (avec et sans prise en compte des graphes décrivant les sens-usages) ont été évalués en utilisant une partie (20 %) du sous-corpus correspondant à chaque mot ambigu, mise de côté dès le début.

La fréquence d'utilisation des EQVs de traduction dans le corpus étant très variable, nous avons pris soin d'inclure dans le corpus d'évaluation des segments correspondant à tous les EQVs. Le nombre de segments compris dans le corpus

de test pour chaque EQV est proportionnel à sa fréquence dans l'ensemble du sous-corpus du mot ambigu. Le corpus de test de *movement* comprend 88 segments au total, celui de *class* 59 et celui de *competence* 55.

4.2. Principes de l'évaluation

Les résultats pouvant être considérés comme corrects sont les cas où :

1. une seule proposition de traduction est faite et est correcte
2. plusieurs propositions sont faites et la première (selon une classification en fonction de leur poids respectif) est la proposition correcte
3. plusieurs propositions sont faites et la proposition correcte n'est pas la première mais une autre dans la liste des résultats.

La justesse de la prédiction de traduction est estimée par rapport à la **traduction de référence** repérée dans le corpus de test, c'est-à-dire l'EQV qui traduit la nouvelle instance du mot ambigu dans le corpus. La méthode est évaluée à l'aide des métriques de **précision** et de **rappel**, où la précision correspond au rapport des prédictions correctes faites sur le nombre total de prédictions faites, tandis que le rappel correspond au rapport des prédictions correctes faites sur le nombre total de cas de référence.

4.3. Présentation des résultats de l'évaluation

Les résultats de l'évaluation du processus de WSD et de prédiction de traduction pour l'ensemble des mots étudiés sont présentés dans le Tableau 6.

	critères d'évaluation	précision	rappel
traits – correspondances inter-langues	(1), (2)	83 %	59 %
	(3)	92 %	66 %
traits – graphes des EQVs	(1), (2)	74 %	71 %
	(3)	94 %	91 %

Tableau 6. Résultats de l'évaluation

Les deux premières lignes du tableau représentent les résultats de l'évaluation lorsque les ensembles de traits décrivant les correspondances inter-langues sont utilisés pour la WSD et la prédiction de traduction. La première ligne concerne les résultats corrects obtenus d'après les critères d'évaluation (1) et (2) : le rappel de 59 % signifie que des prédictions correctes, d'après ces deux critères, sont faites pour 59 % des nouvelles instances, tandis que le taux de précision de 83 % signifie que 83 % des propositions faites sont correctes. La deuxième ligne décrit les résultats obtenus si les cas concernés par le troisième critère d'évaluation sont également considérés comme corrects. Les deux dernières lignes du tableau montrent les résultats obtenus lorsque seuls les graphes des EQVs sont utilisés pour la WSD et la prédiction de traduction. Ces résultats présentent une amélioration évidente par rapport au cas précédent. La troisième et la quatrième lignes du tableau décrivent respectivement les résultats obtenus par application des critères (1) et (2), et (3).

Ces résultats, divergents selon la prise en compte ou non des correspondances établies entre les graphes de cooccurrence des deux langues, s'expliquent par la différence entre les ensembles de traits auxquels le nouveau contexte est comparé. Lorsque les correspondances entre les graphes des deux langues sont utilisées, le contexte des nouvelles instances est comparé aux ensembles de traits qui décrivent ces correspondances. Ces ensembles ne contiennent qu'une sous-partie des traits des graphes mis en relation, c'est-à-dire leurs associations communes (autrement dit, leurs traits communs). D'où la probabilité que les associations du contexte de la nouvelle instance ne soient pas trouvées dans ces ensembles, ce qui justifie le bas rappel.

En revanche, lorsque seuls les graphes des équivalents sont utilisés, les associations du contexte des nouvelles instances sont comparées à la totalité des associations trouvées dans les graphes des EQVs. Ces ensembles d'associations étant beaucoup plus grands que ceux utilisés dans le premier cas, il n'y a que très peu de nouvelles instances pour lesquelles une correspondance n'est pas trouvée.

Nous avons remarqué que les propositions erronées concernent essentiellement les cas d'EQVs très rares, où la quantité de segments qui leur correspondent dans le corpus d'apprentissage est très faible (de 1 à 3). La performance de la méthode dépend donc fortement du nombre d'occurrences repérées pour les EQVs et donc, de la taille du corpus d'entraînement.

5. Bilan de l'utilisation d'une méthode monolingue d'acquisition de sens dans un cadre bilingue

Malgré les bons résultats obtenus lors de l'évaluation du processus de désambiguïsation et de prédiction de traduction, la méthode décrite ci-dessus présente un certain nombre d'inconvénients significatifs.

5.1. Nature des distinctions sémantiques proposées

5.1.1. Nécessité de regroupement des sens

Les correspondances créées de la manière décrite dans le paragraphe 2.4. ne permettent pas d'avoir une idée claire de la sémantique des mots ambigus ni de celle de leurs EQVs de traduction. Les distinctions proposées sont de **granularité trop fine** et pourraient difficilement être acceptées comme décrivant des sens des mots ambigus. Bien évidemment, les limites entre **sens** et **usages**, dans le cadre d'une étude contextuelle, ne sont pas très claires. Cependant, la granularité des distinctions proposées est parfois si petite et leur similarité si grande que nous ne pouvons nier la nécessité d'un **regroupement**. Les distinctions repérées correspondent plutôt à des usages des mots, qui doivent être regroupés afin de dégager des sens.

Pour qu'un tel regroupement ait lieu, il faut que des informations pouvant « lier » les usages soient repérées dans les textes. Une des manières possibles aurait pu consister à utiliser des informations plus abstraites sur des classes auxquelles appartiendraient les mots du contexte. La **classification** des mots est souvent considérée comme diminuant l'effet de la dispersion des données sur les résultats d'une analyse distributionnelle, qui est la principale cause de la granularité très fine des distinctions proposées.

Un autre type d'informations utiles à un tel regroupement concerne le lien des sens de granularité fine à un domaine précis. Les informations de domaine sont en effet exploitées par Magnini et Cavaglià (2000) et Magnini *et al*. (2002) pour le regroupement des sens fins trouvés dans WordNet.

Pourtant, des informations externes permettant une classification des mots des contextes lexicaux ou un regroupement des sens-usages obtenus ne sont pas disponibles dans le cadre de cette expérience. Ainsi, nous avons envisagé une manière alternative pour procéder à ce regroupement.

5.1.2. Regroupement des sens à l'aide des équivalents

5.1.2.1. Fiabilité des équivalents en tant qu'indices de sens

La manière envisagée pour fusionner les sens-usages proposés concerne la prise en compte des EQVs de traduction. Les EQVs étant souvent considérés comme des indices pour l'analyse de la sémantique des mots source, la correspondance entre des sens-usages différents et le même EQV pourrait être considérée comme un indice pour le regroupement des sens-usages en un sens de granularité plus grossière. Ce processus est illustré dans la figure 3.

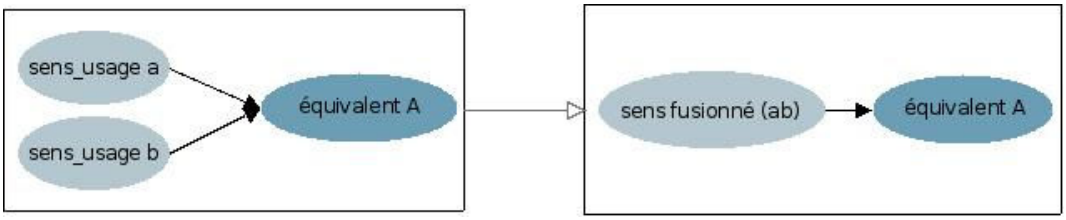


Figure 3. Regroupement de sens-usages en sens à l'aide des EQVs

L'application d'une telle « technique » présupposerait l'existence de **correspondances biunivoques** entre sens et EQVs, qui reposerait sur l'idée qu'un sens est traduit par un EQV. Dans la même ligne de raisonnement, il faudrait considérer la correspondance entre plusieurs EQVs et un sens de la LS comme indice de la granularité trop grossière du sens en question, éventuellement divisible en des sens plus pertinents. Cette division peut, sur la base de la même hypothèse, être effectuée en faisant correspondre un sens distinct à chacun des EQVs de traduction (figure 4).

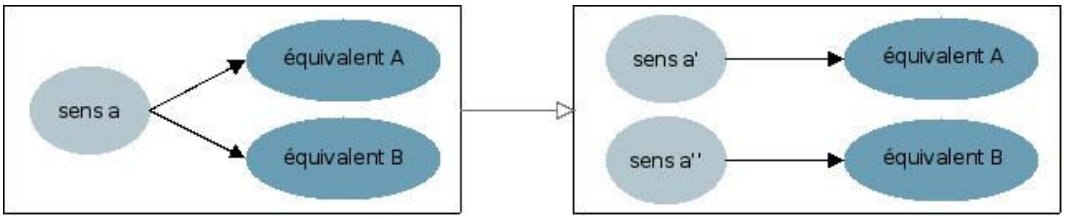


Figure 4. Division d'un sens de granularité grossière en sens plus spécifiques

5.1.2.2. Question de biunivocité entre sens et équivalents

Cependant, la fausseté de l'hypothèse de biunivocité entre sens et EQVs est facile à prouver. En effet, il peut arriver que les mots de deux langues en relation de traduction présentent la même polysémie ou qu'une partie de leur polysémie soit identique[262]. Ainsi, en cas d'existence d'ambiguïtés parallèles entre les langues (phénomène déjà analysé dans le §2.3.2, chapitre 1), l'utilisation des EQVs comme indices de distinctions sémantiques ne s'applique pas.

[262] Les cas où les unités lexicales de deux langues en relation de traduction présentent exactement la même polysémie sont très rares.

En outre, l'adoption d'une telle approche risque de provoquer le **regroupement de sens distants** du mot source, en raison de leur traduction par le même équivalent, qui peut effectivement être ambigu entre les sens en question. Il peut aussi arriver que des EQVs différents correspondent à un sens du mot source sans que cela signifie que le sens en question doive être divisé en plusieurs (sous-)sens. Ceci est le cas, par exemple, d'équivalents qui entretiennent une relation de **similarité sémantique étroite** dans la LC et qui sont éventuellement utilisés pour traduire le même sens du mot source ou des nuances différentes d'un seul sens du mot, ce qui ne justifie pas la proposition de sous-sens distincts. Les relations entre les EQVs pouvant être de nature et de force diverses, ce processus ne permet pas toujours l'induction de distinctions sémantiques pertinentes. Ce risque inhérent à l'utilisation des EQVs de traduction des mots d'une LS comme indices pour l'analyse de leur sémantique a déjà été souligné.

Le recours, de la part du traducteur, à l'utilisation de **synonymes** et de **quasi-synonymes** afin d'éviter la répétition d'éléments lexicaux dans la traduction (souvent lorsque de telles répétitions sont présentes dans le texte source) est considéré comme une des particularités des textes traduits, et même caractérisé comme un **trait universel** (Baker, 1996 : 179). Cette caractéristique des textes traduits peut influencer fortement la réussite des méthodes d'acquisition de sens basées sur les informations de traduction. L'utilisation d'EQVs de traduction différents ne reflétant pas, dans ces cas, des distinctions de sens au niveau de la LS, ces EQVs ne peuvent alors servir à l'analyse de la sémantique des mots ambigus source. Cependant, la distinction entre les cas où l'utilisation d'EQVs de traduction différents reflète des distinctions réelles de sens dans la LS et ceux où elle est due à des raisons liées à la pratique de la traduction n'est pas toujours évidente. Un facteur supplémentaire important posant des problèmes dans le cas d'une analyse de ce type, et qui ne doit donc pas être négligé, concerne la possibilité de **lexicalisation** de distinctions sémantiques au sein de la LC qui ne sont pas effectuées au sein de la LS.

5.1.2.3. Différences de statut des sens proposés

La manière de procéder qui vient d'être décrite peut donc générer davantage de problèmes que ceux qu'elle est censée résoudre. Outre le problème de la **non-biunivocité** entre sens et EQVs, une autre raison de l'inefficacité de cette méthode, dans le cadre de cette expérience, consiste en ce que les distinctions sémantiques proposées par la méthode de désambiguïsation monolingue n'ont pas toutes le même **statut**. Les distinctions reflètent le plus souvent des usages différents du mot ambigu, mais même si des sens pertinents étaient mis en évidence dans les résultats, ils ne pourraient pas être différenciés des usages, ce qui équivaudrait à un traitement similaire[263]. Par conséquent, le regroupement et la division des sens à l'aide des EQVs s'appliqueraient de la même manière à tous les cas, ce qui serait erroné.

5.2. Utilité des informations traductionnelles pour compléter les résultats de la méthode monolingue

5.2.1. Circularité du processus d'identification de sens

Une remarque plus générale qui s'impose est que l'utilisation à ce niveau des EQVs de traduction en tant qu'indices pour le regroupement (ou la division) des sens d'un mot source risque d'initier une sorte de « cercle vicieux ». Rappelons que l'objectif initial de l'utilisation de cette méthode monolingue d'acquisition de sens était l'analyse de la sémantique des mots ambigus de la LS et la création de correspondances sémantiques entre sens et EQVs au sein d'un cadre bilingue. Les correspondances en question devaient permettre

[263] Par exemple, les trois sens-usages proposés pour le mot *class* et décrits par les mots *school*, *number* et *lesson* correspondent en effet à des emplois du mot dans le sens « scolaire », tandis que le sens décrit par *device* reflète le sens de « catégorie », qui pourrait bien être considéré comme un sens distinct du mot.

l'identification des sens du mot source qui auraient pu être véhiculés par chacun de ses EQVs[264].

La prise en compte des EQVs en tant qu'indices pour le raffinement et, même, la correction des résultats de la méthode d'acquisition de sens risque, en fait, de « brouiller » davantage le paysage. Ne disposant pas d'autres informations sur la sémantique des équivalents que celles mises en évidence par le processus de repérage de sens effectué au niveau de la LS – liées aux EQVs par l'établissement des correspondances inter-langues – la considération des EQVs en tant qu'indices pour l'analyse de la sémantique des mots source s'avère très dangereuse : bien que l'un des buts initiaux était de déduire des conclusions sur la sémantique des EQVs par référence aux sens repérés dans la LS, est apparu finalement le besoin de recourir aux EQVs et aux correspondances de traduction établies, pour mieux cerner les sens des mots source.

5.2.2. Calcul de similarité sémantique entre équivalents

La solution que nous avons envisagée à ce problème consiste à utiliser une **mesure de similarité sémantique** qui permet d'aboutir à des conclusions à propos de la sémantique des équivalents. L'estimation de cette similarité rendrait possible la distinction entre les EQVs pouvant servir au regroupement ou à la distinction des sens du mot source, initialement repérés, et ceux qui ne seraient pas utilisables dans ce but. La logique de cette solution est que l'estimation de la similarité des EQVs correspondant au même sens d'un mot source permettrait de juger de la proximité des sens qu'ils véhiculent. Des relations de synonymie ou de quasi-synonymie repérées entre eux pourraient, par exemple, justifier leur mise en correspondance avec le même sens source. Autrement dit, un **fort degré de similarité** entre les EQVs permettrait de juger de l'**homogénéité sémantique** du sens source. En revanche, une **différence sémantique significative** entre EQVs correspondant au même sens pourrait révéler une **distinction sémantique** au sein du sens en question, non repérée lors du processus de repérage de sens ;

[264] Les correspondances ne donnent pas, pour autant, une image complète de la sémantique des équivalents. Nous obtenons seulement une image de leur sémantique liée à celle du mot source qu'ils traduisent.

dans un tel cas, une division du sens serait justifiée. Néanmoins, il ne faut pas exclure dans ce cas, comme nous l'avons souligné plus haut, la possibilité de lexicalisation par les EQVs de sens propres à la LC, dont la projection dans la LS provoquerait une distinction sémantique non pertinente.

Lors de l'implémentation de cette démarche d'estimation de la similarité sémantique des EQVs, nous avons constaté qu'elle pouvait être utilisée à elle seule pour l'acquisition des sens des mots ambigus source et pour l'établissement de correspondances sémantiques inter-langues. En effet, l'utilisation de cette méthode, combinée à la méthode monolingue initialement implémentée, ajoutait un niveau de complexité supplémentaire sans pour autant fournir de solutions à une grande partie des problèmes. Au contraire, ce processus a même constitué, dans certains cas, une source supplémentaire d'erreurs sans fournir de solution à ceux censés être résolus.

Cette constatation nous a fait abandonner la méthode contextuelle monolingue initialement implémentée en faveur d'une deuxième méthode, basée sur le calcul de similarité sémantique entre les équivalents de traduction des mots ambigus. Dans le chapitre suivant, nous présenterons de manière détaillée les principes de fonctionnement de cette méthode ainsi que des exemples d'application aux données de notre corpus, puis nous analyserons une partie des résultats obtenus.

CONCLUSION

Dans ce chapitre, nous avons montré les inconvénients de l'utilisation d'une méthode monolingue d'acquisition de sens dans un cadre de traduction. Les distinctions sémantiques proposées par la méthode implémentée étant de **granularité trop fine**, ces distinctions ne s'avèrent pas pertinentes pour la description des sens véhiculés par les mots source. La finesse des distinctions est en grande partie due à la **dispersion des données** présente dans le corpus, qui ne permet pas d'effectuer de généralisation à partir des usages des mots, ce qui rendrait possible la description de leurs sens.

En outre, la granularité de ces distinctions explique que les correspondances établies entre les mots ambigus et leurs EQVs de traduction ne puissent être caractérisées comme des « correspondances sémantiques ». Ces correspondances mettent tout simplement en évidence les traits du contexte des usages des mots source traduits par chacun des EQVs, ne permettant ainsi pas de tirer des conclusions sur la sémantique des mots source et cible.

Néanmoins, comme il a été démontré par l'évaluation à petite échelle menée ici, les correspondances en question permettent une performance assez bonne du processus de prédiction de traduction. Cette bonne performance paraît possible même si les informations modélisées lors du repérage des sens dans la LS ne sont pas prises en compte, c'est-à-dire en utilisant seulement les graphes de cooccurrence des EQVs.

Ces résultats pourraient même mettre en cause la nécessité du processus de désambiguïsation au niveau de la LS pour la traduction. Mais nous ne nous autorisons pas à émettre des conclusions de ce type, d'une part, en raison de la faible étendue de l'évaluation et, d'autre part, parce que nous souhaiterions obtenir également, hormis de bons pourcentages de rappel et de précision, de bons résultats d'un point de vue **qualitatif**. Nous considérons ainsi qu'un processus de désambiguïsation orienté vers la traduction pourrait fournir des résultats plus satisfaisants et c'est précisément cette piste que nous allons explorer avec la méthode qui sera proposée par la suite.

CHAPITRE 6

ACQUISITION DE SENS DANS UN CADRE BILINGUE

INTRODUCTION

L'absence de lien entre une grande partie des méthodes d'acquisition de sens et de désambiguïsation (WSD) proposées et la finalité des applications où elles peuvent être exploitées constitue, comme nous l'avons déjà souligné, une critique régulièrement adressée à ces méthodes. Plus précisément, ces critiques concernent la non prise en compte des besoins des applications différentes en matière de WSD. Dans un cadre multilingue, par exemple, l'applicabilité de méthodes ne prenant pas en compte les relations d'équivalence entre les unités lexicales de langues différentes peut être facilement mise en doute.

La difficulté à créer des correspondances inter-langues pertinentes au niveau des sens repérés, par une méthode monolingue, a été démontrée dans le

chapitre précédent ; l'implémentation d'une méthode d'acquisition de sens n'utilisant pas d'informations de traduction a été présentée, ainsi qu'une tentative de création de correspondances *a posteriori* entre les résultats obtenus par cette méthode et les informations provenant de la LC. La non-conformité, au sein de certaines applications, de méthodes indépendantes d'un cadre applicatif précis a provoqué l'émergence d'approches plus clairement orientées vers des applications.

Le travail mené ici se situe dans un **cadre de traduction**, aussi l'acquisition de sens et la WSD doivent être sensibles aux particularités de ce cadre et liées à la finalité des applications concernées. Dans la suite de ce chapitre, nous présenterons une méthode d'acquisition de sens dirigée par les données (*data-driven*), où les informations de traduction entrent en jeu dès le début de l'analyse, ce qui permet l'établissement de correspondances sémantiques plus pertinentes entre les mots des deux langues. L'approche proposée appréhende d'une manière originale la question de la polysémie lexicale dans un cadre de traduction. Les résultats de cette méthode sont utilisables pour la WSD et la prédiction de traduction au sein de systèmes automatiques.

1. Acquisition de sens orientée vers la traduction

1.1. Méthode dirigée par les données

La méthode d'acquisition de sens proposée combine des informations provenant des deux langues représentées au sein d'un **corpus parallèle aligné** (bitexte). Plus précisément, il s'agit d'une méthode d'acquisition de sens dirigée par les données, qui combine des informations de **traduction** et de **cooccurrence lexicale**, extraites du bitexte d'apprentissage. Les caractéristiques du corpus d'apprentissage utilisé et les étapes de prétraitement que ce corpus a subies, ont été présentées en détail dans le chapitre 4.

Au niveau des EQVs d'un mot ambigu, cette méthode permet la mise en évidence de leurs relations sémantiques et l'identification des sens qu'ils traduisent. Au niveau du mot ambigu source, elle permet l'analyse de sa

sémantique par **projection inter-langue** d'informations, analyse qui révèle des distinctions sémantiques pertinentes pour la traduction.

1.2. Hypothèses théoriques sous-jacentes

1.2.1. Hypothèses distributionnelles

1.2.1.1. Hypothèse distributionnelle du sens

L'analyse effectuée ici se base sur l'hypothèse contextuelle (ou distributionnelle) du sens (Harris, 1954 ; Firth, 1957a,b), tout comme la méthode présentée dans le chapitre précédent. D'après cette hypothèse, les **sens** d'un mot correspondent à ses **usages** dans les textes, usages qui se reflètent dans le contexte lexical (ou co-texte) qui entoure le mot. Par conséquent, l'étude et l'analyse des contextes lexicaux dans lesquels un mot apparaît permettent d'éclairer sa sémantique.

1.2.1.2. Hypothèse distributionnelle de la similarité sémantique

Une extension de l'hypothèse distributionnelle du sens consiste en l'hypothèse distributionnelle de la similarité sémantique, d'après laquelle des mots sémantiquement similaires apparaissent dans des contextes similaires. Ainsi, la **similarité sémantique** des mots peut être définie comme une **similarité distributionnelle** (Charles et Miller, 1989). Cette hypothèse de la similarité sémantique nous sera très utile par la suite, pour le repérage des sens du mot source par le biais de ses EQVs.

1.2.2. Hypothèse de correspondance sémantique entre mots en relation de traduction

Une autre hypothèse sous-jacente à notre méthode est que les EQVs d'un mot source traduisent ses différents sens dans la LC, qu'il existe donc une sorte

d'**équivalence** (ou de correspondance) **sémantique** entre les unités lexicales des deux langues qui se trouvent en relation de traduction. Au sein d'un corpus textuel, l'équivalence de traduction peut être considérée comme une **équivalence en contexte** (Chesterman, 1998 : 31).

En effet, les contextes source permettent de découvrir les régularités auxquelles les traducteurs ont répondu, consciemment ou pas, en choisissant parmi les équivalents (Mauranen, 2002). Les données de traduction trouvées au sein d'un corpus parallèle peuvent donc être conçues comme le résultat d'un processus, dans lequel le mot traduit a été interprété dans son contexte[1]. Les équivalents traduisant des instances de mots en contexte[2] sont donc considérés véhiculer dans le contexte de la LC à peu près le même sens véhiculé par l'instance du mot source ou, autrement dit, remplir la même fonction. Par conséquent, les régularités des contextes source peuvent servir à décider si les différents équivalents traduisent le même sens du mot source ou non (Mauranen, *ibid*.).

[1] Les traducteurs évaluent les possibilités interprétatives des expressions linguistiques de la LS dans des contextes spécifiques, au sein de textes ayant des objectifs précis, et ils essaient ensuite de récréer les mêmes possibilités d'interprétation dans un texte cible, qui sert un objectif comparable dans une autre langue (Dyvik, 2003, 2005).

[2] C'est-à-dire des usages précis des mots source, dont le sens est accessible en ayant recours aux éléments composant leur contexte.

1.2.3. Hypothèse distributionnelle du sens dans un cadre bilingue

Dans un premier temps, nous acceptons donc les hypothèses suivantes :

1. l'**hypothèse distributionnelle du sens** et de la **similarité sémantique**, selon laquelle le contexte lexical des mots permet l'émission de jugements concernant leur propre sémantique et leur similarité sémantique
2. l'**hypothèse d'une correspondance sémantique inter-langue**, selon laquelle les éléments lexicaux utilisés dans la traduction ont été choisis par le traducteur de manière à ce que le sens du texte traduit soit le plus proche possible du sens de l'original ou, d'un point de vue communicatif, que la fonction remplie par la traduction soit la plus proche possible de la fonction remplie par l'original.

Dans un second temps, nous étendons au niveau des mots de la LC les conséquences de l'acceptation de l'hypothèse contextuelle du sens, en émettant l'hypothèse suivante :

3. les informations venant des contextes lexicaux d'un mot source, lorsqu'il est traduit par un équivalent précis, peuvent éclairer le(s) sens traduit(s) et, par conséquent, véhiculé(s) par cet équivalent.

Comme nous l'avons déjà montré (cf. §2.3, chapitre 1), les correspondances entre les sens d'un mot source et ses EQVs peuvent être plus ou moins complexes. Elles peuvent être **biunivoques**, cas où chaque EQV traduit seulement un sens du mot ambigu en le lexicalisant dans la LC[3]. Elles peuvent aussi être plus complexes, à savoir qu'un EQV peut traduire plus d'un sens ou

[3] Hypothèse sous-jacente aux méthodes d'acquisition de sens qui utilisent les EQVs comme indices de distinctions sémantiques au niveau des mots de la LS, souvent nuancée par la considération de paramètres qui influencent la lexicalisation des sens (cf. §2.1, chapitre 2).

qu'un sens peut être traduit par plus d'un EQV. Avant le repérage des sens du mot source, les correspondances entre le mot source et ses EQVs se situent au niveau lexical. Ce type de correspondance est décrit dans la figure 1.

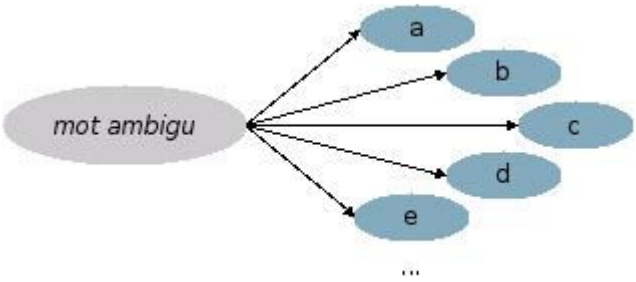


Figure 1. Correspondances entre un mot ambigu et ses EQVs avant le repérage de sens

Etant donné, d'une part, que les sens des mots ambigus source ne sont pas connus à l'avance mais reflétés dans leurs contextes et, d'autre part, qu'il existe des correspondances non évidentes entre ces sens et les EQVs du mot, l'objectif de l'analyse qui sera effectuée est donc de clarifier ces correspondances.

1.3. Similarité distributionnelle : calculée dans quel contexte ?

Dans une méthode monolingue d'acquisition de sens basée sur l'hypothèse distributionnelle, la comparaison des contextes lexicaux des différentes instances d'un mot ambigu permet leur regroupement (clustering) en fonction de leur similarité. Les clusters qui en résultent sont donc considérés refléter les sens véhiculés par le mot (Schütze 1992, 1998). Ici la comparaison ne porte pas sur des contextes « individuels » dans lesquels les instances du mot ambigu apparaissent, mais sur des **ensembles de contextes** de ce mot, constitués relativement à chacun de ses EQVs. Ces ensembles seront donc décrits par référence aux EQVs du mot. Comme nous le montrerons plus loin, pour certaines des expériences menées, les ensembles de contextes ne sont pas formés des contextes du mot source mais des contextes des EQVs dans la LC.

1.3.1. Contexte source des équivalents

Les informations contextuelles relatives à un mot ambigu proviennent du sous-corpus qui lui correspond[4]. Nous appelons **contexte de la LS** (ou **contexte source**) **d'un EQV** du mot ambigu, le contexte lexical des instances du mot source traduites par cet EQV. Ce contexte se trouve dans les **segments de la LS** mis en correspondance avec les **segments de la LC** contenant l'EQV en question, au sein des unités de traduction du sous-corpus constitué pour le mot source[5]. Le repérage de ces segments, dans lesquels le mot source est traduit par chacun de ses EQVs, est fait de la manière décrite dans le paragraphe 2.3.11 du chapitre 4.

Si un EQV traduit plusieurs instances du mot source (au sein d'unités de traduction différentes), le contexte source de l'EQV correspond au contexte du mot source dans l'**ensemble des segments source** (cf. figure 2). Cette figure montre des unités de traduction constituées de segments de la LS et de la LC. Les instances d'un mot source *m* traduites par un EQV précis (*a*) peuvent être décrites comme des instances '*m_a*' (de même pour les instances traduites par d'autres EQVs (*b*, *c*, etc.) : '*m_b*', '*m_c*', etc.). L'EQV peut donc servir à étiqueter les instances en question.

Le contexte source de l'EQV *a* du mot *m* est constitué de l'ensemble des traits retenus du contexte des instances '*m_a*' au sein des segments source (LS). Ces traits sont les **noms**, les **adjectifs** et les **verbes** qui apparaissent avec une fréquence supérieure à 1 au sein des contextes[6]. Le contexte source de l'EQV *a* du mot *m*, dans l'ensemble des unités de traduction où *m* est traduit par *a*, sera décrit comme '***contexte_LS_m_a***' ('*contexte_LS_m_b*', '*contexte_LS_m_c*', etc., pour les autres EQVs).

[4] La manière dont ce sous-corpus est construit est décrite dans le paragraphe 2.3.10, chapitre 4.

[5] Comme nous l'avons montré, les unités de traduction ont été définies pendant la phase d'alignement du corpus au niveau des phrases et correspondent à des segments des deux langues liés par une relation de traduction, qui contiennent de 0 à 2 phrases par langue.

[6] Nous avons constitué une courte liste de mots à filtrer (*stoplist*), qui consiste en des mots de contenu apparaissant dans les 50 premières positions de la liste lemmatisée de fréquence des mots générée à partir du BNC (British National Corpus) (Kilgarriff, 1997a). Notre liste contient les mots suivants : *be*, *have*, *do*, *say*, *go*, *get*, *make*.

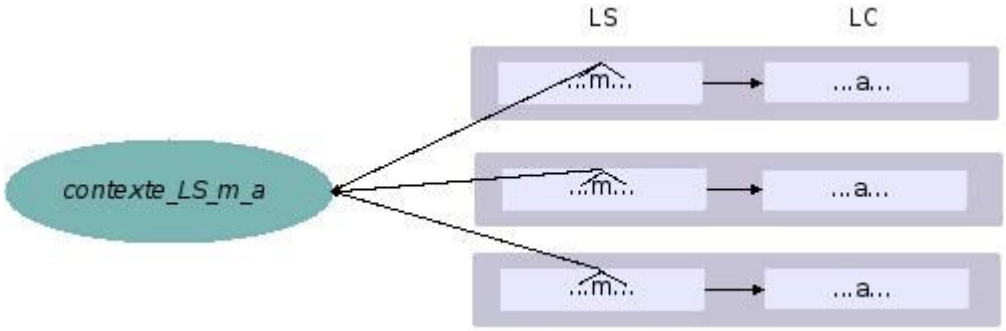


Figure 2. Contexte de la LS d'un EQV du mot ambigu *m*

Les **traits des contextes source** utilisés pour cette estimation de similarité distributionnelle sont les cooccurrents des instances du mot source traduites par chacun des EQVs ('*m_a*', '*m_b*', etc.) et inclus dans les '*contexte_LS_m_a*', '*contexte_LS_m_b*', etc. Les ensembles de contextes constitués de cette manière dans la LS pour chacun des EQVs du mot ambigu, font donc l'objet d'une comparaison qui permet, d'une part, d'estimer la similarité des instances impliquées du mot source et, d'autre part, de juger de la similarité sémantique des EQVs (cf. 5.3.3.).

1.3.2. Contexte cible des équivalents

Outre les contextes source d'un **EQV** du mot ambigu, il est également possible de parler de son **contexte cible** (ou **contexte de la LC**). Ce contexte entoure l'EQV au sein des segments de la LC où il traduit le mot source. Si l'EQV traduit plusieurs instances de ce mot au sein d'unités de traduction différentes, son contexte cible correspond à ses cooccurrents dans l'ensemble des segments de la LC au sein de ces unités, comme cela est décrit dans la figure 3. Le contexte cible de l'EQV *a* du mot *m* dans l'ensemble des unités de traduction où *a* traduit *m* est décrit comme '*contexte_LC_m_a*'.

Les ensembles de contextes constitués dans la LC pour chacun des EQVs du mot ambigu feront aussi l'objet d'un **calcul**, au niveau de la LC, dans le but d'estimer la **similarité distributionnelle des EQVs**. Les traits des contextes de la

LC utilisés pour cette comparaison sont les cooccurrents des EQVs (*a*, *b*, etc.), inclus dans les '*contexte_LC_m_a*', '*contexte_LC_m_b*', etc.

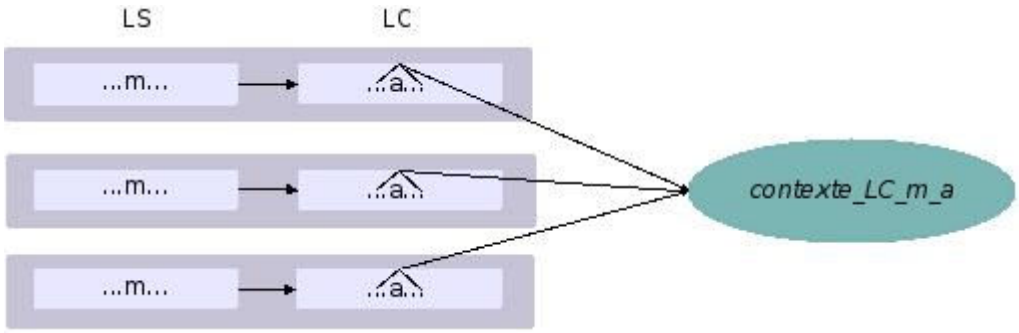


Figure 3. Contexte de la LC d'un EQV du mot ambigu *m*

1.3.3. Distinction des contextes en fonction de la direction de traduction

Initialement, nous avons opéré une distinction supplémentaire à propos des informations du corpus d'apprentissage qui ont été exploitées. Comme nous l'avons déjà souligné (cf. §1.1), la direction de traduction au sein du corpus n'est pas uniforme. En raison du petit volume de textes originaux grecs, une répartition du corpus en deux parties équilibrées en fonction de la langue des originaux n'a pas été possible. Nous avons néanmoins essayé d'explorer les éventuelles différences au niveau des résultats obtenus dans les deux cas suivants :

- lorsque les informations de l'**ensemble du corpus** sont exploitées et
- lorsque l'analyse porte seulement sur la partie du corpus constituée des **textes originaux en anglais** et de leurs traductions en grec.

Plus précisément, dans le cas d'un mot ambigu anglais *m*, entretenant des relations de traduction avec un ensemble d'EQVs grecs (*a*, *b*, *c*, etc.), le calcul de similarité porte sur :

a. le *contexte_LS_m_EQV* de chaque EQV (*a*, *b*, *c*, etc.) dans l'ensemble du corpus

b. le *contexte_LC_m_EQV* de chaque EQV (*a*, *b*, *c*, etc.) dans l'ensemble du corpus

c. le *contexte_LS_m_EQV* de chaque EQV (*a*, *b*, *c*, etc.) dans le sous-corpus constitué par les textes originaux anglais et leurs traductions en grec

d. le *contexte_LC_m_EQV* de chaque EQV (*a*, *b*, *c*, etc.) dans le sous-corpus constitué par les textes originaux anglais et leurs traductions en grec.

Cette distinction, à propos des informations du corpus d'apprentissage exploitées, a vite été abandonnée en raison des divergences observées, qui consistaient en une **détérioration** des résultats dans le cas du **corpus plus restreint**. Les relations de similarité révélées par le calcul de similarité à partir de l'ensemble du corpus se sont généralement avérées plus pertinentes.

Après avoir présenté les principes du processus d'induction de sens, ainsi que les informations du corpus exploitées, le paragraphe suivant sera consacré à une description plus détaillée de l'étape d'apprentissage de notre méthode d'acquisition de sens.

2. Apprentissage automatique pour l'acquisition de sens dans un cadre bilingue

2.1. Acquisition de sens par apprentissage non supervisé

2.1.1. Notions centrales de l'apprentissage supervisé

D'après une définition générale de l'**apprentissage automatique** (*machine learning*) (Mitchell, 1997 : 2), un programme est censé apprendre d'une expérience *E*, par rapport à un ensemble de tâches *T* et une mesure de performance *P*, si sa

performance dans les tâches *T*, mesurée par *P*, s'améliore avec l'expérience *E* [7]. Par conséquent, selon Mitchell, pour qu'un problème d'apprentissage soit bien défini, trois paramètres doivent être identifiés : l'ensemble des **tâches**, la **mesure** de la performance à améliorer et la **source d'expérience**. Cependant, ces paramètres concernent le cas de l'apprentissage **supervisé**. La source d'expérience peut être, dans ce cas, un **corpus d'entraînement** qui contient des **observations** (ou objets ou occurrences) annotées, c'est-à-dire **classifiées**. A l'aide de l'algorithme d'apprentissage, la machine construit un **modèle** (ou fonction cible) pour la tâche en question, qui associe à toute observation *x* une classe *c*. L'apprentissage de la fonction cible permet par la suite à la machine d'effectuer des tâches avec une performance améliorée (par ex. des décisions de classification pertinentes).

2.1.2. Notions centrales de l'apprentissage non supervisé

Contrairement aux algorithmes supervisés, qui utilisent des corpus d'entraînement contenant des observations annotées, les algorithmes employés pour l'apprentissage **non supervisé** se basent sur des **occurrences non classifiées** (ou non étiquetées). Ne nécessitant pas de classification prédéfinie, ces algorithmes repèrent les similarités entre les occurrences présentes dans les données et révèlent leurs **motifs d'association** qui servent, par la suite, à leur **regroupement** (clustering). Un **clustering** est jugé satisfaisant si des groupes d'occurrences « homogènes » sont obtenus en sortie de la méthode – c'est-à-dire des groupes formés d'occurrences les plus similaires possibles.

L'**entrée** de l'algorithme est donc constituée d'**occurrences non classifiées** et la **sortie** de **groupes d'occurrences similaires**. Les techniques d'apprentissage non supervisé et les algorithmes utilisés pour l'acquisition de sens ont été présentés en détail dans le paragraphe 2.1 du chapitre 3. La technique principale d'apprentissage utilisée dans ce but est le **clustering** (ou clusterisation), qui

[7] Dans un problème de reconnaissance d'écriture manuelle, par exemple, la tâche consisterait à reconnaître et à classifier des mots écrits à la main au sein d'images, la mesure de performance *P* pourrait correspondre au pourcentage de mots correctement classifiés et l'expérience d'apprentissage (ou d'entraînement) *E* pourrait consister en une base de données de mots écrits à la main avec des classifications précises.

consiste en la répartition d'un ensemble de données (ou jeu de données) dans des sous-ensembles (clusters), de telle manière que les données du même sous-ensemble partagent certains traits communs. L'objectif est donc l'obtention d'une **bonne partition** du jeu de données.

L'estimation de la proximité des données (objets) repose sur une **mesure de distance** définie à l'avance. Les méthodes traditionnelles de clustering (méthodes de partitionnement et méthodes hiérarchiques) procèdent au regroupement des objets en les comparant deux à deux à l'aide de ce type de mesure (Beck, 2006). La notion de distance est centrale en clustering ; le principe général consiste en effet à minimiser la distance entre deux objets d'un même cluster et à maximiser la distance entre deux objets de clusters distincts. Dans le cas de l'acquisition de sens monolingue, les **objets à clustériser** sont le plus souvent des **instances de mots ambigus** et le clustering s'effectue par prise en compte d'informations provenant de leurs contextes (Schütze, 1992, 1998 ; Pedersen et Bruce, 1997a ; Pantel et Lin, 2002, 2003). Ces informations contextuelles constituent les traits caractérisant les instances et permettent l'estimation de leur similarité (ou de leur distance). Les clusters obtenus dans un tel cadre sont considérés décrire les différents sens des mots ambigus.

2.2. Acquisition de sens non supervisée basée sur des informations de traduction

2.2.1. Clustering au niveau de la LC basé sur des informations source

2.2.1.1. Clustering de quel type d'éléments ?

La méthode d'acquisition de sens proposée dans ce travail est une méthode d'apprentissage automatique **non supervisé** ; plus précisément, il s'agit d'une **méthode de clustering**. L'apprentissage s'effectue en utilisant des informations provenant du **corpus d'entraînement** (appelé aussi corpus d'apprentissage). Cependant, les objets à clustériser ne sont pas les instances d'un mot ambigu en fonction de ses contextes individuels (correspondants à ses différentes instances) – comme c'est le cas dans un calcul de similarité sémantique monolingue – mais

ses EQVs de traduction. Ces EQVs sont trouvés dans le sous-corpus constitué pour le mot ambigu source à partir du corpus d'apprentissage. En outre, contrairement à une méthode de clustering monolingue – où les objets à clustériser et les informations utilisées pour l'induction de sens proviennent de la même langue – ici le **clustering** s'effectue au sein de la **LC** en exploitant des **informations contextuelles** provenant de la **LS**[8].

2.2.1.2. Informations contextuelles exploitées pour le clustering

Les informations contextuelles utilisées pour le clustering proviennent donc de la LS et, plus précisément, des ensembles de contextes constitués pour chacun des EQVs du mot ambigu. Ces **ensembles de contextes** font l'objet d'une série de comparaisons par paire, processus qui permet d'estimer leur similarité. Les traits des contextes source utilisés pour cette comparaison sont les cooccurrents des instances du mot source traduites par chacun des EQVs et inclus dans les ensembles '*contexte_LS_mot-ambigu_EQV*' ('*contexte_LS_m_a*', '*contexte_LS_m_b*', etc.).

2.2.1.3. Conclusions déduites de la similarité des contextes source

Bien évidemment, la comparaison des ensembles de contextes constitués du côté de la LS pour un mot ambigu montre également le degré de similarité des contextes du mot ambigu : un **fort degré de similarité** indique l'**homogénéité sémantique** du mot source, tandis qu'une **dissimilarité** constitue un indice des **distinctions sémantiques** caractérisant le mot.

Sur la base de la troisième hypothèse du paragraphe 2.2.3[9], la comparaison des ensembles de contextes source correspondant aux EQVs permet aussi d'estimer le degré de **similarité sémantique** des EQVs. Une forte similarité des contextes source indique l'existence de relations sémantiques entre les EQVs et sert à leur **regroupement**. En revanche, la dissimilarité des ensembles de

[8] Comme nous le montrerons plus loin, la possibilité d'exploiter des informations contextuelles provenant de la LC a été aussi explorée, mais nous avons finalement opté pour l'utilisation des informations contextuelles source.

[9] Pour rappel, cette hypothèse découle de l'hypothèse distributionnelle monolingue et de l'hypothèse d'existence d'une correspondance d'ordre sémantique entre le mot source et ses EQVS.

contextes source comparés est considérée comme montrant la dissimilarité sémantique des EQVs, qui traduisent des sens éventuellement différents du mot ambigu. Ainsi, le **clustering** se base sur une comparaison des ensembles de contextes du mot source constitués pour chacun de ses EQVs :

> Deux équivalents sont clustérisés si la similarité de leurs ensembles de contextes source respectifs s'avère suffisamment forte.

La mesure permettant d'estimer cette similarité ainsi que le seuil qui sert à juger de la pertinence de la relation entre deux EQVs sont présentés en détail plus loin (cf. §2.3).

Cette méthode d'acquisition de sens repose sur l'hypothèse que la notion de similarité sémantique, basée sur la similarité distributionnelle, peut être reprise ici afin d'émettre des jugements sur la similarité sémantique des EQVs des mots ambigus. Les EQVs traduisant des mots en contexte – c'est-à-dire des usages précis des mots dont le sens est accessible par recours aux éléments composant le contexte – véhiculent dans le contexte de la LC un sens similaire à celui de l'occurrence du mot source ou, autrement dit, remplissent la même fonction.

2.2.1.4. Projection inter-langue des clusters de sens

Les résultats de l'estimation de la proximité sémantique des EQVs sur la base des régularités observées dans les contextes source, modélisés en termes de **clusters**, peuvent être **projetés** de nouveau sur le **mot source**. La projection du clustering effectué sur le mot ambigu source permet l'analyse de sa sémantique. Un cluster, constitué de la manière décrite ci-dessus, est supposé désigner un sens du mot source. Les sens de ce mot peuvent être plus ou moins distingués (ou liés), selon les relations existantes entre les clusters correspondants.

2.2.1.5. Récapitulatif des principes sous-jacents au clustering

Pour récapituler : sur la base des trois hypothèses présentées dans le paragraphe 2.2.3 et après comparaison, à l'aide du calcul de similarité, des ensembles de contextes source correspondant à deux EQVs, le clustering s'effectue de la manière suivante :

a. si la **similarité** entre les ensembles de contextes source correspondant aux EQVs est **forte**, les EQVs sont alors **clustérisés**. Le clustering se fonde sur la considération que la similarité sémantique des EQVs est également élevée ou, autrement dit, qu'ils traduisent le même sens du mot source.

b. si les ensembles de contextes source correspondant aux EQVs s'avèrent dissimilaires, les EQVs sont alors **distingués**. Cette distinction se fonde sur la considération que la similarité sémantique des EQVs traduisant ces instances **n'est pas** non plus **élevée** ou, autrement dit, qu'ils traduisent des **sens différents** du mot source. Dans ce cas, les EQVs appartiennent à des clusters distincts.

Nous parvenons, de cette manière, à obtenir une image de la sémantique des EQVs utilisés, tout en tenant compte du contexte lexical des instances du mot source qu'ils traduisent. Cette méthode se différencie nettement de celle initialement utilisée (et présentée dans le chapitre précédent), dans le sens où nous ne nous appuyons pas sur les résultats d'un processus de repérage de sens au sein de la LS, effectué sans prise en compte des EQVs, pour créer des correspondances inter-langues. En revanche, les informations de distribution lexicale provenant de la LS sont combinées aux informations de traduction (c'est-à-dire, les informations sur les EQVs des mots ambigus) **dès le début** du processus d'analyse sémantique. Cette combinaison permet d'établir des correspondances inter-langues pertinentes et d'aboutir à des conclusions concernant la sémantique des mots des deux langues.

2.2.2. Clustering de vecteurs de traductions

La méthode de clustering utilisée ici est à différencier d'autres méthodes utilisées dans un cadre bi-(ou multi-)lingue pour l'acquisition de sens, dans lesquelles le contexte lexical d'un mot ambigu est conçu uniquement dans le sens traductionnel et correspond à ses traductions possibles dans un ensemble de langues. Ide *et al*. (2001, 2002) proposent une telle méthode (cf. §2.2.2, chapitre 2).

Dans cette méthode, les sens véhiculés par un mot ambigu source sont également révélés à l'aide d'un processus de clustering s'appliquant à ses instances au sein des textes d'un bitexte multilingue, constitué de six versions parallèles d'un texte anglais (LS). Le clustering, dans ce cas, se base sur les traductions des instances du mot dans l'ensemble des langues cibles. Un vecteur est construit pour chaque instance du mot ambigu, qui représente ses traductions possibles dans les six langues. Au début du processus, chaque vecteur forme un cluster à lui seul ; l'ensemble des clusters d'un mot ambigu constitue l'entrée d'un algorithme d'agglomération, qui fusionne les paires de clusters de manière itérative, sur la base de la distance minimale entre eux. Les clusters qui résultent de ce processus représentent les différents sens et sous-sens du mot ambigu.

Une **différence** importante entre cette approche et celle que nous proposons se situe au niveau de la conception du **contexte lexical** qui, dans le cas de la méthode d'Ide *et al*., est constitué des **traductions des mots ambigus** dans un ensemble de langues, tandis qu'ici, il concerne **le contexte source des EQVs du mot ambigu**.

2.2.3. Clustering au niveau de la LC basé sur des informations cible

Une autre piste que nous avons explorée consiste à calculer la similarité des EQVs en prenant en compte leur propre **contexte lexical** dans la **LC**. Selon notre deuxième hypothèse de départ, les EQVs d'un mot ambigu, au sein d'un bitexte, traduisent les différents sens du mot dans la LC. Ainsi, de la même manière que le calcul de similarité des EQVs par référence aux contextes source peut aboutir à des clusters d'EQVs reflétant les sens du mot ambigu, le clustering des EQVs sur

la base des contextes au sein desquels ils véhiculent ces sens, pourrait servir aussi à mettre en évidence les sens en question.

Lorsque la comparaison a lieu dans la LC, les informations contenues dans les unités de traduction correspondant au mot source, et regroupées en fonction de ses EQVs, sont également utilisées. Néanmoins, dans ce cas, les traits contextuels utilisés pour l'estimation des similarités sont les cooccurrents de chaque EQV trouvés dans l'ensemble des contextes cible qui lui correspond et qui est décrit comme '*contexte_LC_mot-ambigu_EQV*' ('*contexte_LC_m_a*', '*contexte_LC_m_b*', etc.). La constitution de l'ensemble des contextes cible correspondant à un EQV se fait de la manière illustrée dans la figure 3 (cf. §1.3.2.). Cette approche ressemble davantage aux tâches d'estimation de la similarité sémantique entre mots dans un cadre monolingue (dans le but, par exemple, de création de thésaurus sémantiques (Grefenstette, 1992)). Cependant, les contextes de la LS s'avèrent plus utiles pour l'estimation de la similarité et le clustering des EQVs d'un mot ambigu (cf. §3 de ce chapitre).

Le **rapprochement** de deux EQVs d'un mot ambigu s'effectue donc en fonction de leur proximité sémantique, évaluée par les résultats du calcul de similarité distributionnelle portant sur leurs ensembles de **contextes source** respectifs. Ce calcul constitue la **mesure de distance** utilisée pour le clustering, qui détermine la manière dont la similarité entre les objets est estimée. Le clustering sémantique des EQVs les plus similaires (ou proches) d'après les résultats du calcul, s'effectue par le **programme SEMCLU**, présenté dans la section 3.5. Dans le paragraphe suivant, nous allons exposer le fonctionnement et l'implémentation du calcul de similarité appliqué à nos données, dont les résultats rendent possible le clustering des EQVs.

2.3. Calcul de similarité sémantique

2.3.1. Mesure de similarité

Pour que l'algorithme de clustering puisse répartir les données (objets) dans des clusters dont les éléments partagent des traits communs, une **mesure de distance** est nécessaire pour évaluer leur **proximité**. L'estimation de cette proximité permet ensuite le rapprochement ou la distinction des objets. Cette

mesure de distance est appelée aussi **mesure de similarité** ou de **dissimilarité**, en fonction du cadre dans lequel le clustering s'applique.

La mesure utilisée dans le cadre de ce travail est un **calcul de similarité sémantique**, basé sur le principe de la **similarité distributionnelle**. Ce calcul concerne des paires d'objets, plus précisément d'EQVs d'un mot source, et porte sur les ensembles de contextes qui leur correspondent dans les deux langues (cf. §1.3).

La comparaison des ensembles de contextes correspondants aux EQVs permet leur regroupement ou leur distinction : le regroupement des EQVs signifie qu'ils entretiennent entre eux une relation de similarité sémantique forte, induite de leur similarité distributionnelle élevée, et qu'ils traduisent ainsi le même sens du mot ambigu. En revanche, leur distinction signale leur distance sémantique, induite de leur dissimilarité distributionnelle, et le fait qu'ils traduisent des sens différents du mot source.

2.3.2. Mesure de Jaccard pondérée

2.3.2.1. Dans un cadre monolingue

La mesure utilisée pour estimer la similarité sémantique des EQVs est la **version pondérée** de la mesure de Jaccard. La mesure binaire de Jaccard calcule la valeur de similarité entre deux mots en comparant les traits qu'ils partagent et ceux qui les distinguent. La mesure de Jaccard pondérée considère, en outre, un poids global et un poids local pour chaque trait : le **poids global** prend en compte le nombre de mots auxquels un trait est associé, tandis que le **poids local** concerne la fréquence d'occurrence d'un trait avec un mot précis.

Cette mesure a été déjà utilisée dans un contexte monolingue pour la création de thésaurus sémantiques. Le système de Grefenstette (1994 : 48-50) par exemple, qui utilise cette mesure, repère des relations de similarité sémantique entre les mots en se basant sur des contextes syntaxiques de granularité fine[10]. Dans ce cadre, la similarité des mots est calculée sur la base de leurs traits syntaxiques communs et le résultat consiste en des listes de mots similaires.

2.3.2.2. Dans un cadre bilingue

Cette méthode d'estimation de la similarité sémantique des mots est reprise ici dans un cadre bilingue dans le but d'estimer la similarité des EQVs des mots ambigus d'une LS. Dans ce cas, les traits exploités pour évaluer la similarité des EQVs ne sont pas syntaxiques ; il s'agit d'informations d'association des mots et concernent leur **cooccurrence** dans des contextes bien précis. En outre, les traits ne proviennent pas toujours de la langue des éléments à clustériser ; comme nous l'avons déjà montré, les traits des EQVs des mots ambigus peuvent provenir des contextes source qui leur correspondent. En effet, des expériences d'estimation de la similarité sémantique des EQVs ont été menées sur la base de traits provenant des deux langues (LS et LC). Cependant, comme nous l'avons

[10] Par exemple, chaque nom trouvé dans le texte est considéré comme un objet et les mots qui le modifient sont considérés comme ses traits : un nom peut être modifié par un adjectif, un autre nom ou un verbe, et les modifications possibles sont considérées comme les traits connus du nom.

également souligné, les informations provenant de la LS se sont avérées de meilleure qualité et, ainsi, le clustering s'est basé sur ces informations.

Par la suite, nous présenterons les résultats de l'application de ce calcul dans les deux langues et expliquerons pourquoi nous avons finalement choisi d'utiliser les informations de la LS.

2.3.3. Calcul de similarité : sur quelles données ?

Le calcul de similarité des paires d'EQVS s'effectue deux fois pour chaque mot source étudié :

a. par prise en compte des contextes source des EQVs
b. par prise en compte des contextes cible des EQVs.

Les informations contextuelles proviennent du sous-corpus constitué pour le mot ambigu au sein du corpus d'apprentissage, filtré en fonction de ses EQVs (cf. §§2.3.10 et 2.3.11, chapitre 4). Les contextes source et cible d'un EQV *a* sont formés de la manière décrite dans le paragraphe 1.3. Les traits des **contextes source** sur lesquels porte le calcul de similarité sont les **mots de contenu** (plus précisément, les noms, les adjectifs et les verbes) qui apparaissent avec une fréquence supérieure à 1 au sein des *contextes_LS_m_a*. Une **liste de fréquence** est donc construite pour ces traits, à partir de cet ensemble de contextes. Cette liste montre la fréquence de cooccurrence des traits retenus avec le mot *m* au sein des contextes de la LS lorsqu'il est traduit par l'EQV *a* (de même pour les autres EQVs à partir des ensembles *contexte_LS_m_b*, *contexte_LS_m_c*, etc.).

En revanche, dans le cas des **contextes cible** d'un EQV *a*, les traits sur lesquels porte le calcul de similarité sont les noms, les adjectifs et les verbes qui apparaissent avec une fréquence supérieure à 1 au sein du *contexte_LC_m_a*. De la même manière, une **liste de fréquence** des traits est construite à partir de cet ensemble de contextes. Cette liste montre la fréquence de cooccurrence des traits retenus avec l'EQV *a* au sein des contextes cible (de même, pour les autres EQVs à partir des *contexte_LC_m_b*, *contexte_LC_m_c*, etc.). Ce processus est décrit schématiquement dans la figure 4, pour deux équivalents (*a*, *b*) d'un mot *m*.

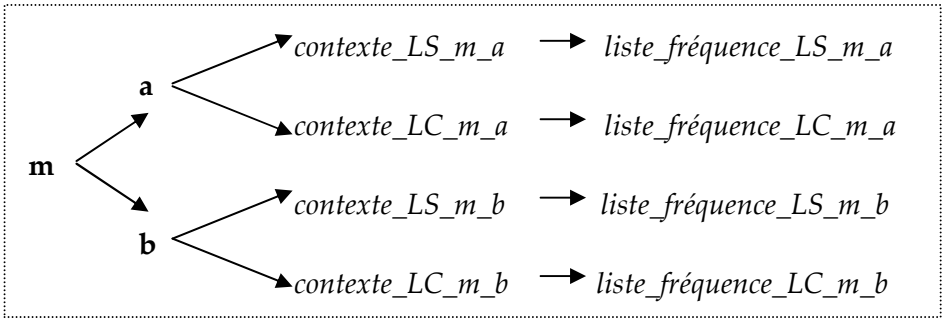


Figure 4. Construction de listes de fréquence source et cible

2.3.4. Définitions formelles

L'**entrée** du calcul de similarité de deux EQVs, dans la LS ou dans la LC, est constituée de la liste de fréquence construite pour chacun d'eux, ainsi que des listes de fréquence construites pour les autres EQVs du mot ambigu dans la langue concernée. Pour chaque **trait (*j*)** du contexte d'un **équivalent (*i*)**, son **poids global** (*pg*), son **poids local** (*pl*) et son **poids total** (*p*) sont calculés à l'aide des formules suivantes :

$$pg(trait_j) = 1 - \frac{\sum_{i=1}^{nbri} p_{ij} \log(p_{ij})}{nrels}$$

Figure 5. Poids global du trait *j* (formule 2.3.4.1)

$$p_{ij} = \frac{fréquence\ absolue\ du\ trait_j\ avec\ l'\ EQV_i}{nombre\ total\ de\ traits\ pour\ l'\ EQV_i}$$

Figure 6. Probabilité d'occurrence du trait *j* avec l'EQV *i* (formule 2.3.4.2)

$$nrels = nombre\ total\ de\ relations\ extraites\ du\ corpus\ pour\ le\ trait_j$$

Figure 7. Nombre d'EQVs avec lesquels le trait *j* est lié (formule 2.3.4.3)

$$pl(EQV_i, trait_j) = \log(\textit{fréquence du trait}_j \textit{ avec l' } EQV_i)$$

Figure 8. Poids local du trait *j* avec l'EQV *i* (formule 2.3.4.4)

$$p = pg * pl$$

Figure 9. Poids total du trait *j* par rapport à l'EQV *i* (formule 2.3.4.5)

Les mots du contexte qui nous servent de traits (*j*) sont pondérés en fonction de leur **dispersion** au sein du sous-corpus constitué pour le mot ambigu (poids global) et de leur **fréquence d'occurrence avec chaque EQV** précis *i* (poids local). Le **poids global** d'un trait (formule 2.3.4.1) dépend du nombre total d'EQVs auxquels il est lié (au sein du sous-corpus du mot ambigu) (formule 2.3.4.3) et de sa probabilité d'occurrence avec chacun des EQVs (formule 2.3.4.2).

Le **poids local** d'un trait avec un EQV précis est calculé sur la base de sa fréquence d'apparition avec l'EQV en question (formule 2.3.4.4). Si un trait apparaît plus d'une fois avec un EQV, le poids total de cette association (formule 2.3.4.5) est donné en multipliant le poids global du trait avec le logarithme de sa fréquence d'apparition avec l'EQV en question (poids local). Ainsi si deux EQVs ont un trait en commun mais que ce trait apparaît plus fréquemment avec l'un qu'avec l'autre, son poids sera plus élevé pour l'EQV avec lequel il apparaît le plus fréquemment. Le logarithme est utilisé pour atténuer l'effet, sur le calcul de Jaccard, des traits qui apparaissent souvent. Plus petit est le nombre d'associations communes entre deux EQVs, plus faible est leur score de similarité.

Le coefficient de Jaccard pondéré (JP) pour deux EQVs *m* et *n* est calculé par la formule 2.3.4.6 :

$$JP(EQV_m, EQV_n) = \frac{\sum_{1}^{nbrj} \min(p(EQV_m, trait_j)p(EQV_n, trait_j))}{\sum_{1}^{nbrj} \max(p(EQV_m, trait_j)p(EQV_n, trait_j))}$$

Figure 10. Jaccard pondéré de deux équivalents (formule 2.3.4.6.)

L'avantage à utiliser le coefficient de Jaccard pondéré (par différence avec le coefficient de Jaccard binaire) est de permettre d'attribuer une **importance** aux traits du contexte par rapport à chaque EQV. La pondération des traits est en effet proportionnelle à leur **pertinence** pour l'estimation de la similarité des mots.

Le numérateur de la mesure de Jaccard prend en compte les traits qui sont **communs** aux deux EQVs : lorsqu'un trait est lié à seulement l'un des deux EQVs, son poids minimal est alors égal à 0 (ce qui correspond à son poids par rapport à l'EQV auquel il n'est pas lié). Ainsi, un trait a un poids supérieur à 0 dans le numérateur uniquement lorsqu'il caractérise les deux EQVs. Le dénominateur de la mesure de Jaccard prend en considération l'ensemble des traits qui caractérisent chacun des EQVs, c'est-à-dire leurs traits **uniques** aussi : si un trait caractérise les deux EQVs, son poids maximal est retenu, tandis que, dans le cas où il ne caractérise que l'un des deux EQVs, c'est son poids avec cet EQV qui est retenu et additionné (ce poids constitue évidemment le poids maximal, étant donné que son poids par rapport à l'EQV auquel il n'est pas lié est égal à 0).

2.3.5. Similarité : proportionnelle au nombre de traits communs ?

Une remarque qui s'impose est que le mot *x* estimé par le Jaccard comme étant le plus similaire à un mot *y* n'est pas celui avec lequel *y* partage le plus grand nombre de traits (Grefenstette 1994 : 52). Il est possible que d'autres mots partagent davantage de traits avec lui, mais qu'ils soient considérés comme

moins similaires. Tel est le cas lorsque les mots apparaissent très fréquemment dans le corpus : le nombre de traits distinctifs entre les mots comparés – qui apparaît au dénominateur de la mesure de Jaccard – est alors très élevé, diminuant, par là, le résultat du calcul de similarité.

2.3.6. Problème dans le cas des mots de basse fréquence

Mis à part ses avantages, un inconvénient important de la mesure de Jaccard pondérée concerne les résultats relativement irrationnels fournis pour des mots peu fréquents dans le corpus. Soit deux mots dont le premier est caractérisé par cinq traits et le deuxième par trois et partagent deux traits en commun, leur coefficient de Jaccard est alors irrationnellement élevé. La non pertinence de la similarité révélée entre ces mots est due à la **petite quantité d'informations contextuelles** disponibles. Cependant, lorsque les mots apparaissent fréquemment (avec une fréquence supérieure à 10) dans le corpus, les résultats obtenus sont satisfaisants.

2.3.7. Exploitation des résultats du calcul de similarité pour le clustering

Le calcul de similarité présenté ici révèle les relations de similarité sémantique entretenues entre les EQVs d'un mot ambigu, relations qui sont exploitées, par la suite, par l'**algorithme de clustering sémantique**. La similarité sémantique des EQVs correspond à leur similarité distributionnelle, révélée par l'analyse de leurs contextes respectifs.

Outre le fait de permettre l'attribution d'un score à une paire d'EQVs – qui correspond à une **quantification** de leur similarité – cette analyse met également en évidence des aspects intéressants à propos de la sémantique des mots comparés. Un tel aspect, qui sera très utile lors de l'exploitation des résultats du clustering sémantique pour la WSD et la prédiction de traduction pour de nouvelles instances des mots ambigus, concerne les traits qui rapprochent et qui distinguent les EQVs. La nature et l'utilité de ces ensembles de traits (autrement appelés **contextes assimilateurs** et **dissimilateurs**) seront décrites plus loin (cf. chapitre 7).

Après avoir défini la mesure de similarité qui sera utilisée pour le clustering, dans le paragraphe suivant, nous allons présenter le fonctionnement du processus de clustering, c'est-à-dire la manière dont le regroupement des mots s'effectue en exploitant les résultats du calcul de similarité.

2.4. Clustering sémantique par une approche de programmation dynamique

2.4.1. Problème global et sous-problèmes

Le **problème global** qui doit être résolu par la méthode de clustering proposée est la construction de **clusters d'EQVs similaires**, clusters dont le nombre n'est pas connu *a priori*. Ce problème global peut pourtant être conçu comme composé d'un ensemble de **sous-problèmes**, dont chacun concerne l'estimation de la **similarité** entre une **paire d'EQVs**. La solution au problème global peut donc être fournie en utilisant les solutions trouvées pour les sous-problèmes qui le composent. En effet, les solutions possibles de clustering sont nombreuses, mais une seule constitue la solution optimale ; le clustering peut donc être exprimé en termes de **problème d'optimisation combinatoire**. Selon le **principe de l'optimalité** (Bellman, 1957), la solution optimale d'un problème de taille n s'exprime en fonction de la solution (optimale) de problèmes de taille inférieure à n (sous-problèmes).

Une solution naïve aux problèmes d'optimisation de ce type consisterait à énumérer puis à évaluer toutes les partitions (ou alternatives de clustering) possibles et à ne retenir que la meilleure d'entre elles (solution optimale). Une **implémentation récursive naïve** peut pourtant provoquer un gaspillage de temps, par le recalcul de solutions à des problèmes déjà résolus.

2.4.2. Résolution du problème global par programmation dynamique

Une technique bien adaptée à la résolution de problèmes d'optimisation combinatoire, où un ensemble de choix doivent être opérés pour obtenir une

solution optimale globale et qui sont composés de sous-problèmes se chevauchant, est la **programmation dynamique** (Bellman, 1957). Cette technique s'avère efficace pour la résolution de ce type de problèmes combinatoires, où un sous-problème donné peut réapparaître jusqu'à obtenir la solution optimale, en évitant de calculer plusieurs fois la même chose. Elle consiste précisément à **mémoriser** des solutions aux problèmes déjà résolus (résultats intermédiaires) et à **réutiliser** ces solutions, si le besoin de résoudre le même problème ressurgit, afin d'aboutir à de nouveaux résultats. Dans une telle approche, la solution optimale à un problème est recherchée en divisant le problème en sous-problèmes et en combinant les solutions partielles, selon le principe « diviser pour régner », qui est son principe de base. Dans le cas du clustering, un algorithme de programmation dynamique converge, à la fin du processus, vers la meilleure solution de clustering.

Les **résultats intermédiaires**, dans le cas de la méthode de **clustering** proposée ici, concernent la **proximité** entre les **EQVs** formant une **paire** ; cette proximité est décrite par le score attribué à la paire par le calcul de similarité. Les résultats fournis par le calcul sont stockés dans une **table** : chaque clé de la table contient une paire d'EQVs dont la similarité est calculée (sous-problème), tandis que la valeur de la clé correspond au score attribué à la paire. Cette table contient les solutions aux sous-problèmes qui composent le **problème global**, à savoir la définition des **clusters finaux**.

La solution au problème global peut donc être atteinte en utilisant les solutions des sous-problèmes qui le composent ; autrement dit, la construction des clusters finaux s'effectue en se basant sur la similarité des paires d'EQVs. Ainsi, en mémorisant les solutions aux problèmes moins complexes, la solution au problème général peut être trouvée par simple **consultation** de la table. La méthode de programmation dynamique implémentée ici est **ascendante** (*bottom-up*), ce qui signifie que tous les sous-problèmes pouvant paraître utiles sont d'abord résolus, et que leurs solutions sont ensuite exploitées pour résoudre le problème global.

Dans la mesure où le clustering effectué vise à regrouper des EQVs similaires, les résultats fournis par le calcul de similarité concernant les paires d'EQVs seront utilisés pendant **toutes** les étapes de clustering. Ces résultats

intermédiaires n'ont donc pas à être recalculés à chaque itération de l'algorithme ; ils sont mémorisés une fois pour toutes (à l'intérieur de la table de similarité) et peuvent ainsi être réutilisés pour trouver de nouveaux résultats.

L'approche adoptée est une approche tabulaire : à chaque appel de l'algorithme de clustering, nous regardons dans la table si la valeur dont nous avons besoin est déjà calculée. Si oui (ce qui est vrai pour toutes les paires possibles d'EQVs), nous ne la recalculons pas mais récupérons la valeur mémorisée. La programmation dynamique permet ici d'éviter les calculs redondants (qui sont implicites dans une approche d'énumération totale). La figure 11 illustre le rôle joué par la table de similarité. Le processus de clustering adopté est décrit en détail dans le paragraphe suivant.

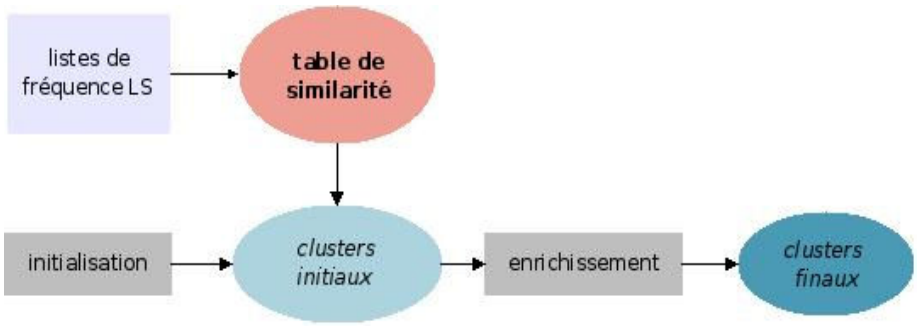


Figure 11. Utilisation de la table de similarité lors du processus de clustering

2.5. Clustering sémantique : le programme SEMCLU

2.5.1. Processus de clustering

L'objectif du clustering sémantique effectué par notre programme, appelé **SEMCLU** (SEMantic CLUstering), est de produire un **ensemble de clusters** d'EQVs pour **chaque mot ambigu**. Les résultats obtenus à partir du calcul de similarité (cf. §2.3) des paires d'EQVs du mot ambigu sont utilisés pour initier le processus de clusterisation.

Ce processus consiste en deux étapes : une première étape visant à produire les ensembles de **clusters initiaux** à deux éléments ; une deuxième étape visant à **enrichir** ces clusters par des éléments supplémentaires.

2.5.1.1. Processus de construction des clusters initiaux

Au préalable, chacun des EQVs est considéré comme membre unique d'un cluster. Etant donné que les résultats du **calcul de similarité** révèlent les relations de proximité entre les EQVs, ce calcul constitue la **mesure de distance** qui conditionnera leur regroupement. Deux EQVs peuvent donc être regroupés en fonction de leur score de similarité évalué par rapport à un seuil. Ce seuil est défini localement pour chaque mot ambigu et correspond à la moyenne des scores de similarité attribués aux paires d'EQVs de ce mot qui sont supérieurs à 0. La comparaison entre le score de similarité d'une paire d'EQVs et le seuil définit la **pertinence** de la relation de similarité correspondante.

Plus précisément, en considérant les scores de similarité attribués aux paires d'EQVs, deux EQVs présentant une similarité suffisamment élevée (c'est-à-dire ayant un score supérieur au seuil) sont considérés comme formant un **cluster à deux éléments**. La formation de ces clusters de deux éléments découle directement du calcul de la table de similarité. Les clusters formés par des EQVs entretenant une relation de similarité **pertinente** (dont le score est supérieur au seuil) sont alors considérés comme de **bons clusters**. En revanche, les paires d'EQVs n'entretenant pas de relations de similarité forte sont incluses dans une liste de **mauvais clusters**. Tous les clusters résultant de cette première étape de clustering, effectuée en utilisant les résultats du calcul de similarité, contiennent donc deux éléments.

2.5.1.2. Processus de construction des clusters finaux

L'**entrée** de la deuxième étape de clustering du programme SEMCLU est constituée des **bons** et des **mauvais clusters**, construits lors de l'étape précédente. Lors de cette étape, les possibilités d'enrichissement des bons clusters de deux éléments sont examinées ; les mauvais clusters sont stockés et ne seront utilisés

qu'à la fin du processus de clustering, afin de créer les clusters à un élément qui feront partie des clusters finaux. L'enrichissement éventuel d'un cluster par de nouveaux éléments dépend des relations de similarité entre ces éléments et les éléments présents dans le cluster. La **sortie** de l'algorithme propose l'**ensemble de clusters générés** pour le mot ambigu et représente la solution optimale de clustering de ses EQVs.

L'algorithme de clustering procède de la manière suivante :

1. L'algorithme **prend en entrée** :
   - la **liste des EQVs** d'un mot ambigu, trouvés dans le lexique bilingue
   - la **table de similarité** du mot (dans la LS)
   - le **seuil**, qui correspond à la moyenne des scores attribués aux paires des EQVs du mot, sans considération des paires avec score égal à 0.
2. La liste des EQVs, la table de similarité et le seuil constituent l'entrée de la fonction ***création_clusters_initiaux***, qui génère les '**bons'** et les '**mauvais clusters'** de deux éléments. Cette fonction renvoie également la partie de la table de similarité du mot M qui contient les paires ayant un score supérieur au seuil (moyenne).
3. La liste de 'bons clusters' issue de *création_clusters_initiaux* contient des doublons (étant donné que la relation entre chaque paire d'EQVs est trouvée deux fois dans la table, par ex. table[EQV1][EQV2] et table [EQV2][EQV1]). Les doublons sont éliminés afin de ne retenir chaque cluster qu'une seule fois.
4. Les mots inclus dans de 'bons clusters' sont mis dans une liste plate, sans doublons ('liste_EQVs_clustérisés').
5. Chaque **bon cluster** (Cluster_C) constitue l'entrée de la fonction ***enrichissement_cluster***, qui entreprend de l'enrichir par d'autres éléments. Pour ce faire, elle se sert de la liste de tous les mots contenus dans de bons clusters (listeEQVs_clustérisés). La fonction *enrichissement_cluster* renvoie chaque cluster éventuellement enrichi par des EQVs ayant une relation significative avec ceux déjà contenus dans le cluster d'entrée. A noter que, étant donné que le cluster d'entrée contient

uniquement deux EQVs, chaque EQV candidat pour être inclus dans le cluster est comparé **uniquement** à ces deux EQVs. Dans le cas donc où d'autres EQVs ont été inclus, par la fonction *enrichissement _cluster*, dans le cluster avant lui, il se peut que ces EQVs n'aient pas de relation significative avec lui (l'existence d'une telle relation n'est pas vérifiée). Le cluster généré par la fonction *enrichissement_cluster* est donc considéré comme un **cluster temporaire**. Autre élément important : l'enrichissement des 'bons clusters' peut résulter à leur fusion (cf. §2.5.3) ou à leur chevauchement (cf. §2.5.4).

6. Chaque cluster temporaire issu de la fonction *enrichissement_cluster* constitue l'entrée de la fonction ***purge_cluster***. Cette fonction entreprend, quant à elle, de supprimer du cluster temporaire les éléments n'entretenant pas de relations avec **TOUS** les autres éléments du cluster, sachant que tous les éléments d'un cluster final doivent être liés de manière significative entre eux. Lorsque cette condition est satisfaite et qu'il ne reste plus d'éléments dans le cluster qui ne sont pas liés avec tous les autres, *purge* s'arrête.
7. La liste des clusters enrichis et nettoyés peut contenir des doublons ; ces doublons sont éliminés pour ne retenir chaque cluster qu'une seule fois dans la liste.
8. La prochaine étape consiste à créer une liste des éléments des 'mauvais clusters' ('listeM_Cm') qui n'ont pas été inclus dans AUCUN des clusters enrichis et nettoyés. La liste plate des éléments inclus dans les clusters en question coïncide avec la 'liste_EQVs_clustérisés' (qui contient l'ensemble des éléments des 'bons clusters' initialement créés). Pour trouver les éléments de la 'listeM_Cm' qui n'ont pas été inclus dans aucun des clusters précédemment générés, nous calculons la **complémentaire** de 'liste_EQVs_clustérisés' dans 'listeM_Cm' (notée 'C_{listeM_Cm}liste_EQVs_clustérisés') ; la complémentaire correspond ici à l'ensemble formé des éléments de 'listeM_Cm' qui **ne sont pas** dans 'liste_EQVs_clustérisés'. Autrement dit, il s'agit de l'ensemble des éléments de 'listeM_Cm' qui lui sont uniques, c'est-à-dire qui

n'appartiennent pas à son intersection avec la 'liste_EQVs_clustérisés' (si les deux ensembles ont des éléments en commun).

9. Chaque EQV de la liste générée lors de l'étape précédente est, ensuite, inclus dans un cluster ne contenant QUE cet EQV ; les clusters qui apparaissent plus d'une fois dans cette liste sont éliminés.
10. La liste '**clusters_finaux'** du mot ambigu (M) contient, d'une part, les clusters enrichis et nettoyés (clusters_enrichis_nettoyés) et, d'autre part, les clusters générés pendant l'étape 9, dont chacun ne contient qu'un seul EQV.

L'algorithme de clustering utilisé est décrit plus en détail en Annexe B. Nous estimons néanmoins important d'analyser davantage quelques principes sous-jacents au fonctionnement du processus de clustering.

2.5.2. Condition d'arrêt : connectivité globale au sein des clusters

Le regroupement des objets ou des clusters dans les méthodes de clustering s'arrête :

- si **un seul cluster** subsiste (c'est-à-dire lorsque tous les clusters ont été fusionnés, ce qui survient à la dernière étape des algorithmes d'agglomération)
- si **chaque objet** est l'**unique membre** d'un cluster (dernière étape des algorithmes de partitionnement)
- si une autre **condition d'arrêt** est vérifiée.

Cette condition d'arrêt peut concerner, par exemple, l'obtention d'un nombre *k* de clusters (où *k* est le nombre de clusters souhaité, si ce nombre est connu à l'avance)[11]. Dans notre cas, ce nombre n'est pas connu ; les clusters d'EQVs sont supposés décrire les sens du mot ambigu source, dont le nombre n'est pas connu à l'avance. Les sens en question sont révélés à partir des données.

[11] Tel est le cas de l'algorithme '*k*-means' qui permet la répartition d'un ensemble de données en *k* classes.

Ainsi, il est souhaité que le clustering s'arrête lorsque l'ensemble des sens véhiculés par le mot source sont repérés et décrits par les clusters générés.

Le clustering effectué par notre algorithme s'arrête lorsqu'une **solution de clustering optimale** est trouvée. Dans cette solution optimale :

- tous les EQVs du mot source doivent être inclus dans un cluster
- toutes les relations (pertinentes ou non) entre EQVs doivent avoir été prises en compte
- les éléments des clusters doivent tous être liés entre eux par des relations de similarité pertinentes.

Un cluster de ce type, dont tous les éléments sont liés entre eux, peut être décrit, dans les termes de la théorie des graphes, comme un **graphe complet**. Un graphe complet est un graphe simple[12], où une arête lie chaque paire de nœuds distincts[13], ce qui signifie que toutes les paires de sommets sont adjacentes[14]. Si les relations entre les éléments d'un cluster généré sont décrites de cette manière, le graphe correspondant doit contenir toutes les arêtes possibles[15], comme cela apparaît dans la figure 12.

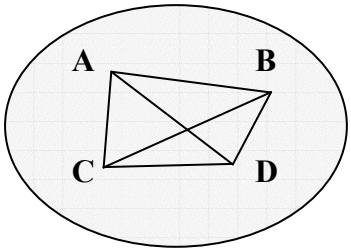


Figure 12. Cluster d'éléments directement liés

[12] Un graphe est dit « simple » s'il ne contient pas de boucles et s'il n'y a pas plus d'une arête reliant deux mêmes sommets.

[13] Le graphe complet de *n* nœuds a *n* nœuds et n(n-1)/2 arêtes, et est dénoté comme k_n. On parle alors de graphe régulier de degré *n-1*.

[14] Deux sommets sont adjacents s'ils sont reliés par une arête. Deux sommets adjacents peuvent aussi être qualifiés de « voisins ».

[15] Les graphes complets sont des cliques d'eux-mêmes. Dans un graphe *G*, une clique est un sous-graphe complet de *G*, c'est-à-dire une partie de l'ensemble des sommets de *G* telle que le graphe induit par *G* sur cette partie soit un graphe complet. Dans la théorie des graphes, une clique est un ensemble de sommets deux à deux adjacents (notion de graphe complet). Le terme « clique » est également souvent utilisé pour parler du graphe induit par une clique.

Le processus de clustering s'arrête lorsqu'il n'est plus possible de générer des clusters en formant un graphe complet (condition d'arrêt).

Cette condition constitue la **condition d'arrêt** de notre algorithme de clustering. Les éléments trouvés dans les clusters finaux contenant plus d'un élément doivent donc **tous** entretenir des relations sémantiques pertinentes directes entre eux. Cette condition est vérifiée lors de l'étape 6, décrite dans le paragraphe 2.5.1.2. Si un cluster temporaire contient un élément qui n'est pas lié à tous les autres éléments du cluster, cet élément est éliminé et n'est ainsi pas inclus dans le cluster final.

2.5.3. Fusion des clusters

Lors du processus d'enrichissement d'un 'bon cluster' initial par de nouveaux éléments, le phénomène de fusion de clusters peut s'observer. Ainsi si, par exemple, il y a une relation pertinente entre les paires d'EQVs 'A-B', 'B-C', 'A-C', 'A-D', 'C-D' et 'B-D', les clusters initiaux construits sont les suivants :

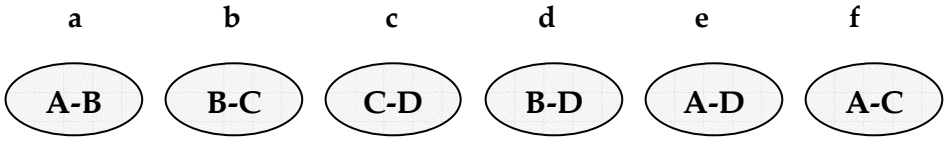


Figure 13. Clusters initiaux

Lors de l'examen du cluster {A, B} dans le but de l'enrichir, les EQVs C et D sont ajoutés au cluster, parce qu'ils ont des relations significatives avec les EQVs qu'il contient. Ainsi, nous obtenons le cluster {A, B, C, D}.

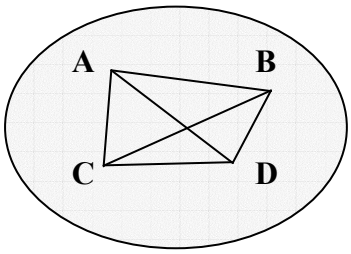


Figure 14. Cluster final

Néanmoins, le cluster {C, D} est également enrichi, à son tour, par A et B, pour la même raison et ainsi de suite pour les autres clusters de la figure 13. A la fin du processus, les « petits » (clusters à deux éléments) clusters ont tous été enrichis par de nouveaux éléments. En éliminant les doublons des clusters créés de cette manière, nous ne retenons, à la fin qu'un cluster qui regroupe tous les éléments. Ainsi, les petits clusters initiaux sont fusionnés au sein du cluster final.

2.5.4. Chevauchement des clusters

Il est possible qu'un élément entretienne des relations pertinentes avec des éléments de **plusieurs clusters**, sans qu'il y ait de relations entre les autres éléments des clusters en question. Dans ce cas, les clusters présentent des recouvrements **sans être fusionnés**, parce que leurs éléments ne sont pas tous liés. Les clusters qui présentent des recouvrements de ce type **se chevauchent**. Ainsi si, par exemple, il existe une relation pertinente entre les paires d'EQVs 'A-B,' 'A-C' et 'D-E', les clusters formés lors de la première étape de clustering, qui génère les clusters initiaux, seront les suivants.

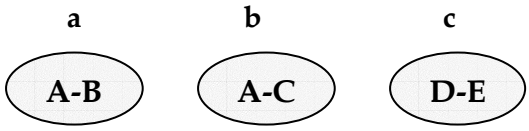


Figure 15. Ensemble de clusters issus de la première étape de clustering

Les clusters *a* et *b* présentent un recouvrement qui concerne l'élément 'A'. Néanmoins, ils ne peuvent être fusionnés car il n'y a pas de relation pertinente entre les éléments 'B' et 'C' ; ainsi, si la fusion avait lieu, les éléments du cluster

résultant ne seraient pas tous liés entre eux, ce qui n'est pas autorisé. Le chevauchement des clusters *a* et *b* pourrait être schématiquement illustré comme suit.

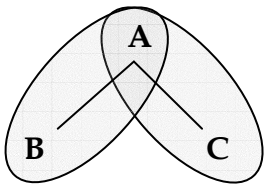


Figure 16. Chevauchement de deux clusters à deux éléments

Au niveau du programme, le chevauchement entre deux clusters (ou plus) est décrit par l'**inclusion des mêmes éléments** dans les clusters en question, c'est-à-dire par l'**intersection non vide** de leur contenu. Il se peut aussi qu'un cluster se chevauche avec plusieurs autres clusters. Dans notre exemple, si une relation pertinente existe entre les éléments 'C' et 'F', le cluster *b* se chevaucherait avec deux clusters, comme décrit dans la figure 17.

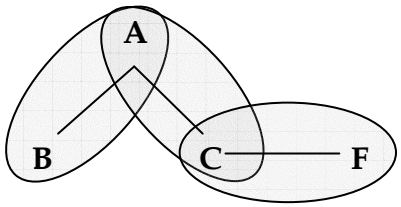


Figure 17. Chevauchement de trois clusters à deux éléments

De tels chevauchements sont également possibles aux étapes suivantes de clustering, où les clusters contiennent éventuellement un plus grand nombre d'éléments. Ainsi, si nous disposons, par exemple, des relations suivantes : 'A-B', 'A-C', 'A-F', 'B-C', 'C-F', les clusters finaux construits sont décrits dans la figure 18.

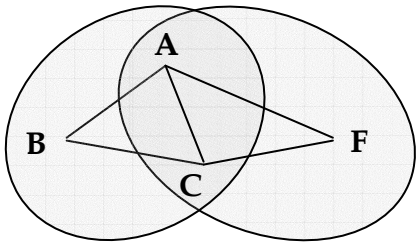


Figure 18. Chevauchement de clusters à plus de deux éléments

Nous reviendrons sur le sujet du chevauchement des clusters dans le paragraphe suivant, où nous présenterons les avantages pouvant être tirés de la possibilité de création de clusters se chevauchant offerte par notre algorithme de clustering.

2.6. Clustering flou pour la représentation des relations inter-sens

2.6.1. Chevauchement sémantique

La méthode de clustering proposée permet la création de **clusters non mutuellement exclusifs**. Il s'agit donc d'une **méthode floue** (*fuzzy*)[16]. Un des avantages de la modélisation de ce type de chevauchements concerne la possibilité de décrire les relations éventuellement entretenues entre les EQVs d'un cluster et des EQVs situés dans d'autres clusters.

Cette modélisation permet ainsi de représenter les relations pouvant exister entre clusters (cf. figures 16-18, §2.5.4). Les clusters d'EQVs étant considérés représenter les sens du mot ambigu, les relations repérées entre eux peuvent être alors interprétées comme des **relations inter-sens**. La description de ces relations constitue un avantage de l'analyse sémantique effectuée.

La non possibilité de représentation des relations inter-sens constitue, à l'inverse, un reproche souvent adressé aux représentations sémantiques traditionnelles, dans lesquelles les sens sont énumérés et représentés de manière uniforme et leurs relations ne sont pas prises en compte.

[16] Ces méthodes se distinguent des méthodes « dures » (*hard*) qui ne permettent pas la création de clusters se chevauchant.

2.6.2. Degré d'appartenance à un cluster

Dans les méthodes de clustering flou, un objet peut appartenir simultanément à plusieurs clusters selon différents **degrés d'appartenance**. Autrement dit, un élément qui se trouve à l'intersection de deux clusters peut être lié à eux plus ou moins fortement. Dans le cas des clusters créés ici, nous pouvons considérer le degré d'appartenance d'un EQV à un cluster comme proportionnel à la force de ses relations avec les autres EQVs inclus dans le cluster. La force des relations en question est quantifiée à l'aide des scores qui leurs sont attribués.

Le degré d'appartenance d'un objet à un cluster peut être pris en compte ou ignoré, ce qui dépend de l'utilisation envisagée pour les résultats du clustering. Si ces degrés sont pris en compte, il est possible d'obtenir une **partition** en **clusters mutuellement exclusifs** à partir d'une **partition floue**, en attribuant chaque objet au cluster auquel il est le plus fortement lié.

2.6.3. Traduction de différents sens par les EQVs situés à l'intersection de clusters

La possibilité de création de clusters se chevauchant constitue un avantage important de la méthode de clustering proposée, étant donné la nature du problème traité, à savoir la création de **correspondances** au niveau du **sens** entre mots des deux langues en relation de traduction. Le recouvrement des clusters permet de décrire les cas où un EQV peut traduire plus d'un sens du mot source dans la LC. Tels sont, par exemple, les cas d'**ambiguïté parallèle**, où l'ambiguïté du mot source est préservée dans la LC.

Le **degré d'appartenance** d'un EQV à un cluster peut être interprété comme la **tendance** d'utilisation de l'EQV en question pour traduire le sens représenté par le cluster. Ainsi, l'appartenance d'un EQV à un cluster à un degré supérieur relativement à celui qui le caractérise par rapport à un autre cluster, peut être interprétée comme une plus grande tendance à traduire par l'EQV le sens représenté par le cluster auquel il est le plus fortement lié ; la possibilité que cet

EQV traduise aussi le sens représenté par l'autre cluster n'est pour autant pas exclue.

L'utilisation dans un cadre de traduction d'un processus qui créerait des clusters mutuellement exclusifs donnerait une solution non naturelle, qui ne permettrait que la mise en correspondance de chaque EQV avec un seul sens du mot source.

2.6.4. Distance inter-sens et nombre d'EQVs communs

Ce type de modélisation, qui permettrait uniquement la création de clusters mutuellement exclusifs, empêcherait la prise en compte des relations éventuellement existantes entre les sens, ce qui assimilerait le résultat de l'acquisition de sens au résultat des méthodes d'énumération de sens (cf. §1.3.1, chapitre 1). Il ne faut pourtant pas exclure la possibilité qu'un EQV traduise dans la LC des sens distants du mot source, étant lui aussi ambigu. Cependant, si deux clusters de la LC représentent deux sens distincts d'un mot source, la probabilité que leur intersection (i.e. chevauchement) contienne plus d'un ou de deux éléments est faible, car cela signifierait que les sens représentés par ces clusters ne sont pas réellement distincts[17].

Il est donc possible de considérer que la **proximité des sens** représentés par les clusters est **proportionnelle** à l'**étendue de leur recouvrement** : les sens seraient d'autant plus proches que leur recouvrement concernerait un plus grand nombre d'EQVs. La différenciation entre sens proches et distants, effectuée de cette manière, peut faciliter une discrimination des sens par rapport à leur statut. Cette différenciation pourrait être exprimée *a posteriori* par la caractérisation des sens en fonction de leur statut (sens homonymes, sens, sous-sens, nuances de sens), ou elle pourrait simplement être décrite de manière implicite par le degré de distinctivité entre les sens. La prise en compte de ce type de discrimination permet également une différenciation au niveau du traitement au sein des applications qui exploiteraient les résultats du clustering.

[17] En effet, il est rare de trouver plusieurs mots pareillement ambigus (entre sens éloignés) au sein d'une langue. D'après Dyvik (1998a, 2003, 2005), l'ambiguïté contrastive est une propriété **accidentelle** et **idiosyncrasique** des mots. Par conséquent, on ne s'attend pas à trouver des instances de la même ambiguïté contrastive dans le cas d'autres mots de la langue ou de mots d'autres langues.

2.7. Projection inter-langue d'informations sémantiques

2.7.1. Repérage de sens

Les objectifs du clustering effectué par la méthode qui vient d'être présentée sont, comme nous l'avons déjà souligné, le repérage de distinctions sémantiques pertinentes pour la traduction au niveau d'un mot ambigu de la LS et la création de correspondances sémantiques entre ce mot et ses EQVs. Les clusters construits pour un mot ambigu regroupent ses EQVs qui entretiennent des relations de similarité sémantique proches. La similarité sémantique des EQVs est estimée à l'aide d'informations venant des contextes des instances du mot source qui leur correspondent et est, par conséquent, proportionnelle à la similarité de ces contextes.

La projection inter-langue des clusters formés par les EQVs sur le mot source permet l'identification des différents sens véhiculés par ce mot au sein de son sous-corpus (cf. §1.3.10, chapitre 4). Si besoin est, les sens induits de cette manière sur le mot ambigu pourraient être décrits à l'aide des EQVs formant les clusters, qui constitueraient ainsi une sorte de définition. Dans une étape ultérieure de traitement, ces ensembles pourraient même constituer des étiquettes sémantiques pour annoter de nouvelles instances des mots ambigus.

2.7.2. Repérage de relations inter-sens

Il faut pourtant souligner que cette projection inter-langue d'informations ne permet pas simplement une énumération des sens du mot source. Un des caractéristiques de l'algorithme de clustering utilisé, à savoir la possibilité de création de clusters se chevauchant, peut avoir également des conséquences bénéfiques au niveau de la description des sens du mot ambigu : le **recouvrement** entre clusters d'EQVs peut être interprété comme représentant des **relations** entre les sens décrits par les clusters en question. La projection des informations de clustering sur le mot source ne permet donc pas simplement le repérage des sens véhiculés par ce mot, mais aussi une description des relations

éventuellement existantes entre les sens, ce qui peut s'avérer très utile pour une description de la sémantique du mot source.

3. Résultats du processus d'acquisition de sens

3.1. Acquisition de sens basée sur le lexique bilingue manuellement généré

Dans ce paragraphe, nous allons illustrer les résultats obtenus par application de la méthode d'acquisition de sens aux données du lexique bilingue anglais-grec construit lors du **repérage manuel de traductions** (cf. §1.3.8, chapitre 4). Nous montrerons la manière dont les relations entre les EQVs mises en évidence par le calcul de similarité sont exploitées pour le clustering et le repérage des sens des mots ambigus. En outre, nous étudierons les conclusions pouvant en être tirées à propos de la sémantique des EQVs.

Les entrées du lexique anglais-grec manuellement généré sont décrites dans le Tableau 1. Nous y donnons le nombre d'unités de traduction composant le sous-corpus d'apprentissage des mots ambigus anglais et leur répartition en fonction de leurs EQVs.

Mot ambigu	EQVs
competence (161)	**αρμοδιότητα (124)**
	ικανότητα (25)
	δικαιοδοσία (3)
	επάρκεια (3)
	δυνατότητα (3)
	κατάρτιση (1)
	δεξιότητα(2)
movement (436)	**κυκλοφορία (318)**
	κίνημα (13)
	διακίνηση (39)
	κίνηση (35)
	μετακίνηση (24)
	κινητικότητα (7)
occupation (88)	κατοχή (21)
	επάγγελμα (44)
	δραστηριότητα (7)
	απασχόληση (7)
	κατάληψη (5)
	ασχολία (4)
plant (188)	εγκατάσταση (21)
	φυτό (112)
	σταθμός (12)
	εργοστάσιο (18)
	μονάδα (22)
	εργαστήριο (3)
	φυτάριο (0)
communication (621)	**ανακοίνωση (365)**
	επικοινωνία (190)
	ενημέρωση (5)
	πληροφορία (10)
	κοινοποίηση (20)
	ανταλλαγή (3)
	αναφορά (21)
	διαβίβαση (4)
	γνωστοποίηση (3)
paper (266)	έγγραφο (33)
	βίβλος (141)
	βιβλίο (65)
	εργασία (5)
	κείμενο (3)
	χαρτί (12)
	δοκιμασία (7)

Mot ambigu	EQVs
preparation (243)	**προετοιμασία (121)**
	κατάρτιση (10)
	εκπόνηση (11)
	παρασκεύασμα (56)
	παρασκευή (10)
	επεξεργασία (6)
	προπαρασκευή (6)
	σκεύασμα (23)
facility (120)	**εγκατάσταση (61)**
	διευκόλυνση (21)
	δυνατότητα (9)
	μέσο (5)
	υπηρεσία (9)
	εξοπλισμός (4)
	υποδομή (8)
	ίδρυμα (3)
treatment (507)	μεταχείριση (88)
	θεραπεία (234)
	αντιμετώπιση (19)
	επεξεργασία (69)
	περίθαλψη (18)
	αγωγή (62)
	καθεστώς (6)
	διαχείριση (3)
	εξέταση (4)
	κατεργασία(2)
	χορήγηση(2)
power (369)	**εξουσία (99)**
	αρμοδιότητα (94)
	ενέργεια (44)
	δύναμη (66)
	ισχύς (43)
	δυνατότητα (8)
	ικανότητα (2)
	δικαιοδοσία (2)
	καθήκον (4)
	ευχέρεια(7)

Tableau 1. Répartition des sous-corpus des mots ambigus par rapport à leurs EQVs

Afin d'illustrer les résultats de la méthode sur les données de traduction manuellement repérées, nous nous servirons de deux exemples de mots ambigus de la LS ayant un grand nombre d'EQVs dans le corpus : *plant* et *movement*. Les résultats fournis pour les autres mots du lexique manuellement construit sont présentés en Annexe D3. Avant de procéder à l'analyse des résultats, quelques remarques générales s'imposent.

3.1.1. Expériences sur différents types de contextes

Le calcul de similarité entre les paires des EQVs d'un mot ambigu a porté, dans une première phase, sur quatre contextes différents. Le contexte sur lequel chaque expérience a porté est décrit dans le Tableau 2. Le sous-corpus est constitué de textes originaux en anglais et de leurs traductions en grec.

Expériences (exp.)	Contextes LS (anglais : AN)		Contextes LC (grec : GR)	
	tout le corpus	sous-corpus	tout le corpus	sous-corpus
1	x			
2		x		
3			x	
4				x

Tableau 2. Contextes utilisés pour les expériences

Cette distinction a vite été abandonnée, dans la mesure où la qualité des résultats obtenus à partir du sous-corpus des textes originaux anglais est souvent inférieure à celle des résultats obtenus sur l'ensemble du corpus. Nous avons également observé que, dans l'ensemble du corpus, les contextes AN donnent de meilleurs résultats que les contextes GR ; en effet, ces derniers attribuent plus souvent des scores élevés à des paires d'EQVs n'entretenant pas de relations pertinentes. Nous avons remarqué que la pertinence des relations repérées (dans la totalité du corpus) dans les contextes des deux langues ou seulement dans les contextes AN est plus élevée que celle des relations repérées seulement dans les

contextes GR. Etant donné, d'une part, les différences qualitatives au niveau des résultats du calcul de similarité en fonction du contexte utilisé et, d'autre part, le besoin de déterminer un seul type de contexte dont les résultats seraient exploités, ensuite, par le processus de clustering, nous n'avons finalement retenu que les résultats fournis par le calcul sur les contextes AN dans l'ensemble du corpus[18].

3.1.2. Impact du paramétrage sur les résultats

Une autre remarque générale qui s'impose est que les résultats du calcul de similarité distributionnelle sont très « sensibles » aux modifications de paramètres concernant la **quantité** des données exploitées. Ceci s'explique par la nature de la méthode, qui est une **méthode totalement dirigée par les données**. Un changement important au niveau des résultats a été remarqué, par exemple, lorsque le sous-corpus des mots ambigus avait été filtré afin d'en éliminer les unités de traduction où les segments de la LC contenaient plus d'un EQV des mots source[19]. L'élimination de ces unités a alors diminué la quantité d'informations contextuelles disponibles, ce qui a eu un effet direct sur les relations de similarité repérées.

Cet effet est plus évident dans le cas d'EQVs rares pour lesquels peu d'informations contextuelles sont disponibles ; dans ces cas, des relations repérées lorsque l'ensemble des sous-corpus était considéré, n'ont pas été proposées après la modification décrite ci-dessus. De telles divergences au niveau des résultats à partir de données différentes sont attendues dans une étude distributionnelle, où la quantité des données est d'une importance cruciale. Nous considérons que ce type d'effet serait, probablement, d'autant moins important que l'étendue du corpus utilisé pour l'apprentissage serait grande.

[18] Les résultats du calcul de similarité pour *plant* et *movement* sur les autres types de contexte seront également présentés, mais ne seront pas exploités.
[19] Pour les raisons de ce choix, cf. §1.3.11, chapitre 4.

3.1.3. Clustering sémantique pour le mot *plant*

3.1.3.1. Equivalents de traduction de plant

Les EQVs grecs de *plant* sont trouvés dans l'entrée de *plant* au sein du lexique bilingue anglais-grec construit par le repérage manuel de traductions. Lors de la construction de cette entrée, seules les instances du nom *plant* ont été retenues (et non celles du verbe). Ces EQVs sont décrits dans le Tableau 3, accompagnés de leur fréquence d'occurrence en tant que traductions de *plant* au sein de son sous-corpus.

Equivalent	Fréquence
φυτό (fyto)	94
μονάδα (monada)	15
εγκατάσταση (egkatastasi)	14
εργοστάσιο (ergostasio)	14
σταθμός (stathmos)	7
εργαστήριο (ergastirio)	3

Tableau 3. Equivalents de *plant* ordonnés par fréquence

3.1.3.2. Calcul de similarité sémantique des EQVs de plant

Pour que le clustering des EQVs soit possible, leur similarité sémantique doit d'abord être calculée. Un score de similarité est attribué à chaque paire d'EQVs relativement à chaque type de contexte, à l'aide de la mesure de Jaccard pondérée (cf. §2.3.2.). Le Tableau 4 décrit les scores de similarité attribués aux paires d'EQVs de *plant*.

Paires d'EQVs	exp.1	exp.3	exp.2	exp.4
μονάδα (monada) - εγκατάσταση (egkatastasi)	**0,209**	**0,177**	**0,192**	**0,184**
σταθμός (stathmos) - εργοστάσιο (ergostasio)	**0,086**	**0,075**	0	0
σταθμός (stathmos) - μονάδα (monada)	**0,048**	0,032	**0,055**	**0,032**
μονάδα (monada) - φυτό (fyto)	0,045	**0,058**	**0,039**	**0,063**
εγκατάσταση (egkatastasi) – φυτό (fyto)	0,043	**0,059**	**0,04**	**0,071**
μονάδα (monada) - εργοστάσιο (ergostasio)	0,028	0,007	0	0,01
σταθμός (stathmos) - εγκατάσταση (egkatastasi)	0,022	0,041	**0,026**	**0,042**
φυτό (fyto) - εργοστάσιο (ergostasio)	0,015	0,005	0	0
εγκατάσταση (egkatastasi) – εργοστάσιο (ergostasio)	0,013	0,007	0,011	0
σταθμός (stathmos) - φυτό (fyto)	0,005	0,008	0,008	0,011
φυτό (fyto) - εργαστήριο (ergastirio)	0	0,002	0,003	0,003
εργοστάσιο (ergostasio) - εργαστήριο (ergastirio)	0	0	0	0
εγκατάσταση (egkatastasi) - εργαστήριο (ergastirio)				
μονάδα (monada) - εργαστήριο (ergastirio)				
εργαστήριο (ergastirio) - σταθμός (stathmos)				
Moyenne	**0,047**	**0,043**	**0,024**	**0,027**

Tableau 4. Table de similarité des EQVs de *plant*

La **relation sémantique** d'une paire d'EQVs est considérée comme **pertinente** si le score qui lui est attribué est **supérieur ou égal au seuil adopté**. Nous rappelons que le seuil correspond à la moyenne des scores supérieurs à 0, qui figure sur la dernière ligne du tableau.

Les paires d'EQVs sont ordonnées selon leur score dans les contextes AN dans l'ensemble du corpus (deuxième colonne, exp. 1). Les relations ayant un score supérieur au seuil dans ce type de contexte sont pertinentes et permettent le regroupement des EQVs impliqués[20]. En revanche, les EQVs *φυτό* (fyto) et *εργαστήριο* (ergastirio) n'entretiennent de relations fortes avec aucun des autres EQVs.

3.1.3.3. Clusters de sens de plant

[20] En revanche, nous observons que des relations non pertinentes reçoivent un score élevé dans les contextes GR et les contextes du sous-corpus (*μονάδα* (monada) – *φυτό* (fyto), *εγκατάσταση* (egkatastasi) – *φυτό* (fyto).

La table de similarité construite pendant l'étape précédente constitue l'entrée du programme **SEMCLU**, qui génère les clusters décrivant les sens du mot ambigu. La sortie du programme SEMCLU pour *plant* est donc constituée de l'ensemble de clusters suivant :

a. {*μονάδα* (monada), *εγκατάσταση* (egkatastasi)}
b. {*σταθμός* (stathmos), *εργοστάσιο* (ergostasio)}
c. {*σταθμός* (stathmos), *μονάδα* (monada)}
d. *φυτό* (fyto)
e. *εργαστήριο* (ergastirio)

Ces clusters sont décrits schématiquement dans la figure 19[21].

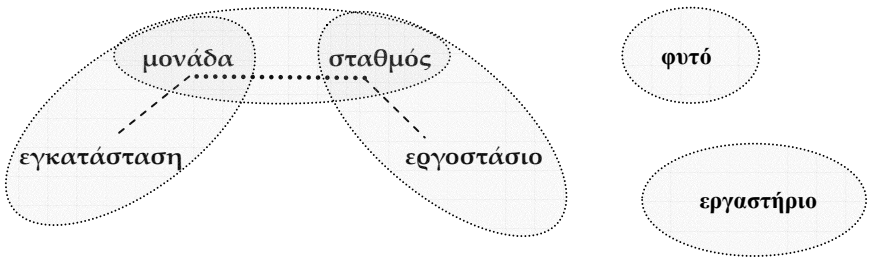


Figure 19. Clusters des EQVs de *plant*

Les trois clusters à deux éléments présentent des recouvrements et s'opposent, de ce fait, aux clusters à un élément, qui se distinguent également entre eux. Etant donné que les clusters d'EQVs construits sont supposés refléter les sens du mot source, leur **distinctivité** pourrait être considérée comme indicative de **distinctions sémantiques** au niveau du mot source. Cependant, est-ce que l'« isolement » d'un EQV par formation d'un cluster à un élément indique toujours une telle distinction ?

[21] Généralement, les clusters contiennent uniquement des éléments, leurs liens étant sous-entendus. Dans ces premiers schémas, nous décrivons pourtant également les liens entre les éléments, afin d'illustrer le type de contexte dans lequel leurs relations ont été repérées. De cette manière, nous distinguons les relations repérées à la fois dans les contextes AN et GR de celles repérées seulement dans les contextes AN à l'aide de formats de lignes différents. Nous rappelons que les relations repérées seulement dans les contextes GR ne sont pas prises en compte pour l'induction des sens et ne sont par conséquent pas décrites dans les schémas.

La **dissimilarité sémantique** de cet EQV par rapport aux autres ne constitue qu'une des raisons pouvant provoquer un tel **isolement**. Une autre raison peut être sa **basse fréquence** d'occurrence : une telle fréquence, couplée au problème de la **dispersion des données**, augmente de manière significative les possibilités d'isolement de l'EQV (que de la dissimilarité sémantique y soit impliquée ou non). Les cas d'EQVs à basse fréquence sont donc assez problématiques à traiter, dans la mesure où la petite quantité d'informations contextuelles disponibles ne permet pas de vérifier l'existence d'une réelle dissimilarité sémantique.

Au contraire, la **haute fréquence** d'un EQV isolé peut constituer un indice beaucoup plus **fiable** pour le repérage d'une réelle distinction sémantique. Dans ces cas, la grande quantité d'informations contextuelles disponibles permet d'exclure la possibilité d'isoler l'EQV par insuffisance d'informations contextuelles. Nous avons testé la validité de ce critère de haute fréquence pour l'estimation de la pertinence sémantique d'une distinction en ajoutant dans le programme un seuil relatif à la fréquence d'occurrence des EQVs (>10). Nous avons effectivement constaté que les clusters à un élément construits en filtrant les EQVs rares illustrent des distinctions sémantiques plus nettes[22].

Dans notre exemple, *φυτό* (fyto) est de loin l'EQV le plus fréquent de *plant*[23]. Le cluster contenant cet EQV est alors considéré refléter un sens distinct de *plant*. Au contraire, l'EQV *εργαστήριο* (ergastirio) traduit seulement 3 occurrences de *plant* ; dans ce cas, il faudrait vérifier si ce cluster décrit une réelle distinction sémantique ou non, vérification qui ne peut avoir lieu seulement sur la base des informations trouvées dans le corpus d'apprentissage. En raison de cette incertitude, le seuil de fréquence d'occurrence des EQVs élimine le cluster en question.

[22] L'adoption de ce seuil de fréquence améliore également, comme nous le montrerons plus loin (chapitre 10), les résultats des méthodes de WSD et de sélection lexicale.
[23] L'EQV *φυτό* (fyto) traduit 64% des occurrences de *plant*.

3.1.3.4. Correspondances sémantiques inter-langues de granularité variable

La distinction entre *φυτό* (fyto) et les EQVs des clusters se chevauchant peut être considérée comme refléter deux **sens non liés** du mot *plant*. Cette distinction de sens est illustrée dans la figure 20. Une **distinction** de **granularité grossière** de ce type coïncide avec le jugement intuitif qui peut être émis sur la sémantique du mot *plant*, à savoir qu'il s'agit d'un cas d'**homonymie**[24], où les unités lexicales impliquées véhiculent deux sens bien distincts : le sens « botanique » et le sens « industriel ».

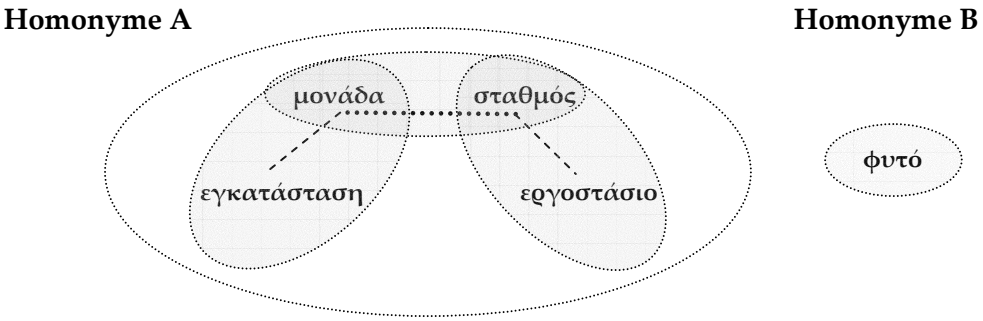


Figure 20. Distinction sémantique de granularité grossière

L'inexistence de relations pertinentes entre certains éléments des clusters se chevauchant (par ex. entre *εγκατάσταση* et *εργοστάσιο* ou *μονάδα* et *εργοστάσιο*) pose des problèmes quant à leur regroupement[25]. Etant donné que chaque recouvrement concerne un seul élément, il est risqué de fusionner les clusters entre eux. Par exemple, dans le cas de l'EQV *μονάδα*, qui se trouve à l'intersection de deux clusters, nous avons les possibilités suivantes :

- les deux clusters décrivent des sens distincts de *plant* entre lesquels *μονάδα* est ambigu

[24] Nous utilisons le terme « homonymie » même si nous admettons que le terme « ambiguïté contrastive » serait plus approprié. L'homonymie implique deux lemmes distincts tandis qu'en travaillant sur un corpus textuel, nous n'avons pas d'informations concernant la distinction entre les lemmes.

[25] Avant l'élimination, du sous-corpus de *plant*, des unités de traduction contenant plus d'un élément, des relations fortes avaient été repérées entre *εγκατάσταση* (egkatastasi) – *σταθμός* (stathmos) et *μονάδα* (monada) – *εργοστάσιο* (ergostasio), tandis que la relation de *εγκατάσταση* (egkatastasi) – *εργοστάσιο* (ergostasio) se trouvait juste au dessous du seuil de pertinence. Il serait intéressant de voir si ces relations auraient été repérées dans un corpus plus grand.

- les deux clusters décrivent des nuances du même sens et la relation entre leurs autres éléments (*εγκατάσταση* et *σταθμός*) n'est pas repérée en raison de la dispersion des données, ce qui empêche la construction d'un cluster plus grand.

Dans la mesure où le non repérage de relations pertinentes entre l'ensemble des éléments de clusters se chevauchant est souvent dû à la dispersion des données, les clusters se chevauchant pourraient éventuellement être regroupés ; le cluster issu de ce regroupement décrirait alors un sens de granularité **grossière**. Le risque, lors d'une telle fusion, de **lier erronément** entre eux des **sens distincts** du mot ambigu augmente d'autant plus que le nombre d'éléments situés à l'intersection des deux clusters est **petit**. Si les clusters se chevauchant de *plant* étaient regroupés, les sens grossiers obtenus pour ce mot, par projection inter-langue des clusters, seraient les suivants :

a. *plant* – {*φυτό*}
b. *plant* – {{*μονάδα*, *σταθμός*}, {*μονάδα*, *εγκατάσταση*}, {*σταθμός*, *εργοστάσιο*}}

Cette manière de représenter le sens décrit par les clusters regroupés (un ensemble d'ensembles d'éléments) permet de décrire le **lien** entre les **sens plus fins** compris au sein d'un des sens homonymiques de *plant*. Mais, un tel regroupement des clusters se chevauchant ne pourrait être effectué qu'*a posteriori*, l'algorithme de clustering du programme SEMCLU ne permettant pas la fusion de clusters contenant des éléments non liés.

Précisons néanmoins que, si de tels sens de granularité grossière étaient formés, ils pourraient tout de même être raffinés par la suite, si besoin était, en dégroupant les clusters fusionnés. Les clusters à deux éléments refléteraient ainsi des distinctions sémantiques plus fines. Dans ce cas, les sens de *plant* seraient les suivants :

a. *plant* – {*φυτό*}
b. *plant* – {*μονάδα* – *εγκατάσταση*}
c. *plant* – {*μονάδα* – *σταθμός*}
d. *plant* – {*σταθμός* – *εργοστάσιο*}

Ces **sens** correspondent en effet à la **solution de clustering** proposée par le programme SEMCLU (cf. figure 17), après élimination du cluster contenant l'EQV *εργαστήριο* (ergastirio) en raison de sa basse fréquence d'occurrence.

Les informations de clustering obtenues peuvent donc être projetées sur le mot source et induire des **distinctions sémantiques de granularité variée**. Les sens mis en évidence au niveau du mot source peuvent être caractérisés (ou même étiquetés) à l'aide des EQVs des clusters permettant leur repérage. Ainsi, les clusters repérés peuvent être utilisés pour créer des correspondances inter-langues entre le mot ambigu et les EQVs traduisant ses sens dans la LC.

3.1.4. Clustering sémantique pour le mot *movement*

3.1.4.1. Equivalents de traduction de movement

Les EQVs de *movement* dans le corpus d'apprentissage et leurs fréquences d'occurrence sont décrits dans le Tableau 5[26].

Equivalent	Fréquence
κυκλοφορία (kykloforia)	251
διακίνηση (diakinisi)	38
κίνηση (kinisi)	28
μετακίνηση (metakinisi)	19
κίνημα (kinima)	11
κινητικότητα (kinitikotita)	6
προσπάθεια (prospatheia)	1
τάση (tasi)	1
βήμα (vima)	1
reste	11

Tableau 5. Fréquence d'occurrence des EQVs de *movement*

[26] La catégorie 'Reste' inclut des cas de changement de catégorie grammaticale, où par exemple le nom *movement* est traduit par un verbe (par ex. …*where the movement of judgments has been fairly free*…/ …*όπου οι δικαστικές αποφάσεις κυκλοφορούν σχετικά εύκολα*…) ou encore des cas où *movement* n'est pas traduit (par ex. …*the ideas of the Biopolitics International Organisation and with the work of the associated international movement*… / …*τις ιδέες της Διεθνούς Οργάνωσης Βιοπολιτικής και τη δραστηριότητα της σε διεθνές επίπεδο*…). Il existe par ailleurs 4 erreurs d'alignement de phrases.

3.1.4.2. Calcul de similarité sémantique des EQVs de movement

Les scores attribués à chaque paire d'EQVs, par le calcul de similarité dans les différents types de contexte, sont présentés dans le Tableau 6.

Paires d'EQVs	exp.1	exp.3	exp.2	exp.4
μετακίνηση (metakinisi) - διακίνηση (diakinisi)	**0,151**	**0,135**	**0,121**	**0,142**
κίνηση (kinisi) - διακίνηση (diakinisi)	**0,131**	**0,144**	**0,107**	**0,161**
μετακίνηση (metakinisi) - κίνηση (kinisi)	**0,11**	**0,174**	**0,108**	**0,205**
κυκλοφορία (kykloforia) - διακίνηση (diakinisi)	**0,098**	**0,098**	**0,073**	**0,099**
κίνηση (kinisi) - κυκλοφορία (kykloforia)	**0,076**	**0,078**	**0,059**	0,069
μετακίνηση (metakinisi) - κινητικότητα (kinitikotita)	**0,073**	0,055	0,05	**0,088**
κινητικότητα (kinitikotita) - διακίνηση (diakinisi)	**0,061**	0,059	0,038	0,058
κίνηση (kinisi) - κίνημα (kinima)	0,047	0,058	0,022	0,051
μετακίνηση (metakinisi) - κυκλοφορία (kykloforia)	0,041	0,05	0,029	0,043
κίνηση (kinisi) - κινητικότητα (kinitikotita)	0,038	0,059	0,007	0,06
κίνημα (kinima) - διακίνηση (diakinisi)	0,027	0,053	0,013	0,04
κυκλοφορία (kykloforia) - κινητικότητα (kinitikotita)	0,02	0,018	0,017	0,019
μετακίνηση (metakinisi) - κίνημα (kinima)	0,02	0,054	0,012	0,07
κίνημα (kinima) - κυκλοφορία (kykloforia)	0,014	0,016	0,008	0,014
κίνημα (kinima) - κινητικότητα (kinitikotita)	0,011	0,055	0	**0,079**
Moyenne	**0,061**	**0,074**	**0,044**	**0,079**

Tableau 6. Table de similarité des EQVs de *movement*

Nous observons que tous les EQVs entretiennent des relations de similarité pertinentes sauf *κίνημα* (kinima). En effet, *κίνημα* traduit le sens abstrait de *movement* (mouvement social, politique, sportif, etc.) et se différencie ainsi des autres EQVs, qui traduisent son sens « physique »[27]. L'EQV *κίνηση* (kinisi) traduit également, dans très peu de cas, le sens abstrait du mot dans le corpus, mais sa relation avec *κίνημα* n'obtient pas un score élevé[28]. Cette relation était pourtant repérée avant l'élimination des unités de traduction contenant plus d'un EQV, c'est-à-dire lorsque le calcul portait sur l'ensemble du sous-corpus de *movement*. Vu l'utilisation très rare de l'EQV *κίνηση* avec ce sens, la modification de la quantité des informations contextuelles disponibles a « affaibli » cette relation[29].

[27] Les autres EQVs peuvent aussi véhiculer le sens abstrait de *movement*, mais pas dans le corpus étudié (par ex. : *διακίνηση (diakinisi)* / *κυκλοφορία (kykloforia)* / *κινητικότητα (kinitikotita) ιδεών* (*movement of ideas*: « circulation d'idées »), *διακίνηση (diakinisi) πληροφοριών* (*movement of information* : « circulation d'informations »).

[28] Par ex. *πολιτική / περιβαλλοντική κίνηση (kinisi)* (mouvement politique, environnemental).

[29] Pourtant, le fait qu'elle était établie dans l'ensemble du sous-corpus rend possible son repérage dans un corpus plus grand. Les EQVs de *movement* qui sont des *hapax* (*προσπάθεια, τάση, βήμα*) et qui ont été exclus du calcul (à cause de leur très basse fréquence d'occurrence), traduisent aussi des sous-sens du sens abstrait de *movement*.

La figure 21 illustre le clustering proposé par SEMCLU pour *movement*.

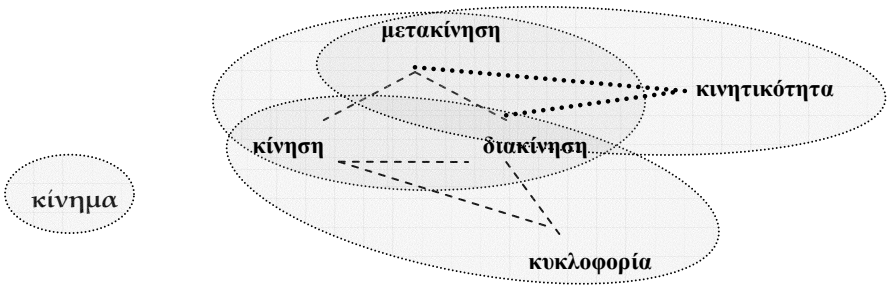


Figure 21. Clusters des EQVs de *movement*

Les EQVs clustérisés traduisent donc le sens « physique » de *movement*. Leurs différences de distribution sont surtout dues au complément de *movement* à ces différentes occurrences (mouvement de qui/de quoi?) traduites par chacun des EQVs : *κυκλοφορία* (kykloforia) est généralement utilisé lorsqu'il est question de la circulation de personnes, de travailleurs, de marchandises, de capitaux, de services et de données ; *διακίνηση* (diakinisi) est surtout utilisé lorsqu'il est question de la circulation de personnes (travailleurs, citoyens), mais aussi de transfert de produits, de marchandises et de capitaux ; *μετακίνηση* (metakinisi) traduit *movement* quand il fait référence au mouvement de personnes et, très rarement, au déplacement d'animaux et de grains (*hapax*) ; *κινητικότητα* (kinitikotita) traduit *movement* quand il s'agit de la mobilité de personnes (citoyens, étudiants). *Κινητικότητα* est plus éloigné sémantiquement des autres mots clustérisés ; il décrit plus spécifiquement la notion de « mobilité », ce qui justifie, d'une part, sa basse fréquence d'utilisation comme traduction de *movement* et, d'autre part, le non repérage de relations pertinentes pour cet EQV dans les contextes grecs[30].

Les **forts recouvrements** entre les trois clusters peuvent donc constituer un indice de **lien** entre les sens qu'ils représentent. Ces recouvrements concernent deux EQVs de chaque cluster et ainsi, le risque d'ambiguïté contrastive des mots

[30] En effet, cet EQV formait un cluster à lui seul, dans les résultats obtenus avant l'élimination des unités de traduction contenant plus d'un EQV.

trouvés dans leur intersection diminue[31]. Le chevauchement des clusters permet donc de représenter les relations inter-sens : les sens décrits par les **clusters se chevauchant** peuvent être considérés comme **moins distincts** et **antagonistes** que celui décrit par le cluster à un élément, qui se distingue des autres.

3.1.4.3. Correspondances sémantiques inter-langues de granularité variable

Les correspondances inter-langues générées pour *movement*, à partir du résultat du processus de clustering (sans regroupement des clusters se chevauchant), sont les suivantes :

a. *movement* – {*μετακίνηση*, *κίνηση*, *διακίνηση*}
b. *movement* – {*κίνηση*, *διακίνηση*, *κυκλοφορία*}
c. *movement* – {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}
d. *movement* – {*κίνημα*}

En prenant en compte le recouvrement des clusters, ces correspondances sont modifiées comme suit :

a. *movement* – {{*μετακίνηση*, *κίνηση*, *διακίνηση*}, {*κίνηση*, *διακίνηση*, *κυκλοφορία*}, {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}}
b. *movement* – {*κίνημα*}

Une solution intermédiaire consisterait à fusionner uniquement les clusters qui présentent le plus « fort » recouvrement. Ainsi, d'après les résultats du calcul de similarité, le cluster {*μετακίνηση*, *κίνηση*, *διακίνηση*} est plus fortement lié au cluster {*κίνηση*, *διακίνηση*, *κυκλοφορία*} qu'au cluster {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}, parce que les relations entre les paires d'EQVs *κυκλοφορία-κίνηση* et *κυκλοφορία-διακίνηση* ont un score plus élevé que celles entre *μετακίνηση-κινητικότητα* et *διακίνηση-κινητικότητα* (cf. table de similarité, §3.1.4.2). Si les clusters les plus proches sont regroupés, nous obtenons des distinctions sémantiques que nous pourrions caractériser comme étant de **granularité intermédiaire** :

[31] Nous avons déjà dit que ce risque est d'autant plus grand que le nombre de mots trouvés à l'intersection des clusters est petit.

a. *movement* – {{*μετακίνηση*, *κίνηση*, *διακίνηση*}, {*κίνηση*, *διακίνηση*, *κυκλοφορία*}}
b. *movement* – {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}
c. *movement* – {*κίνημα*}

Nous avons donc montré comment les résultats du calcul de similarité servent au clustering sémantique des EQVs de *movement*. Les informations de clustering peuvent être alors exploitées pour identifier des **sens de granularité variée** au sein du mot ambigu.

3.2. Acquisition de sens basée sur le lexique automatiquement généré

3.2.1. Remarques générales

Dans ce paragraphe, nous allons présenter un échantillon des résultats obtenus par application de la méthode d'acquisition de sens sur les données du lexique anglais-grec, généré à partir du résultat de l'alignement des mots. Le processus d'acquisition de sens n'a pas été appliqué à la totalité des mots présents dans ce lexique, mais seulement sur un sous-ensemble de **150 entrées** (cf. Annexe D1), construit par application de certains critères. Plus précisément, il s'agit d'entrées ne contenant pas d'EQVs erronés et où le nombre d'EQVs fournis pour le mot source est supérieur à deux. Cette condition correspond à une contrainte de la méthode d'acquisition de sens, qui ne parvient pas à fournir des clusters pertinents dans le cas où il y a seulement deux EQVs ; le seuil déterminant la pertinence d'une relation de similarité étant la moyenne des scores (supérieurs à 0) attribués aux différentes paires d'EQVs, dans le cas où il y a deux EQVs, leur score coïncide avec la moyenne. Par conséquent, les deux EQVs sont toujours clustérisés, même si leur relation n'est pas pertinente.

Des problèmes apparaissent aussi lorsqu'il y a trois EQVs : le seuil étant la moyenne de leurs scores, une des relations entretenues est toujours estimée comme non pertinente, même si les trois EQVs sont tous sémantiquement liés et qu'ils traduisent le même sens du mot source. En général, les résultats de la méthode sont d'autant plus pertinents que le nombre d'EQVs est élevé.

Etant donné que la **précision** (la bonne qualité des EQVs) est privilégiée au sein du lexique automatiquement construit, au détriment du **rappel** (le nombre d'EQVs fournis), les clusters générés à partir de ces données ne concernent pas tous les EQVs qui traduisent les mots ambigus dans le corpus[32] ; les EQVs fournis par ce lexique sont ceux qui ont été repérés par l'alignement lexical et retenus après les filtrages des résultats de l'alignement[33]. Notons que des informations sur la fréquence des EQVs ne sont pas disponibles[34].

Les 150 entrées retenues de ce lexique sont données en Annexe D1, ainsi que les résultats du clustering obtenus pour l'ensemble des mots anglais de ces entrées. Nous présentons ici quelques résultats à titre indicatif.

3.2.2. Clustering sémantique pour le mot *settlement*

Settlement a les EQVs suivants au sein du lexique : *διευθέτηση* (diefthetisi), *διακανονισμός* (diakanonismos), *συμβιβασμός* (simvivasmos) et *οικισμός* (oikismos). Le clustering fourni par SEMCLU pour ce mot ambigu est décrit dans la figure 22.

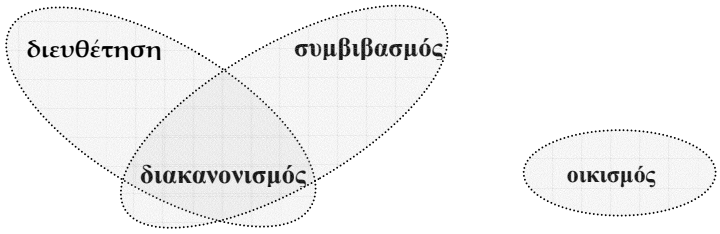


Figure 22. Clusters des EQVs de *settlement*

Les clusters se chevauchant désignent les sens de « règlement » (décrit par le cluster {*διευθέτηση*, *διακανονισμός*}) et d'« arrangement » (décrit par le cluster {*διακανονισμός*, *συμβιβασμός*}), qui sont des sens proches de *settlement*. Ces deux sens se distinguent plus nettement de celui décrit par le cluster à un élément, le

[32] Des 6 EQVs de *movement*, par exemple, seuls 4 sont contenus dans le lexique : *μετακίνηση*, *διακίνηση*, *κίνηση* et *κυκλοφορία*.
[33] Pour une description de ces étapes de filtrage, cf. §2.3.4 et §2.3.5 du chapitre 4.
[34] L'absence de telles informations ne permet pas d'éliminer des résultats les clusters à un élément qui contiennent un EQV de basse fréquence.

sens de « colonie ». Ainsi, les sens retenus pour *settlement* pourraient constituer soit deux sens de granularité grossière (par fusion des clusters chevauchants) :

a. *settlement* – {*διευθέτηση, διακανονισμός*}, {*διακανονισμός, συμβιβασμός*}}
b. *settlement – οικισμός*

soit trois sens de granularité plus fine (ce qui correspond à la sortie de SEMCLU) :

a. *settlement* – {*διευθέτηση, διακανονισμός*}
b. *settlement* – {*διακανονισμός, συμβιβασμός*}
c. *settlement – οικισμός*

3.2.3. Clustering sémantique pour le mot *contribution*

Les EQVs du mot ambigu *contribution* fournis dans le lexique sont les suivants : *συμμετοχή* (symetochi), *συνεισφορά* (syneisfora), *συμβολή* (symvoli), *εισφορά* (eisfora) et *εισήγηση* (eisigisi). Les clusters construits pour ces EQVs sont décrits dans la figure 23.

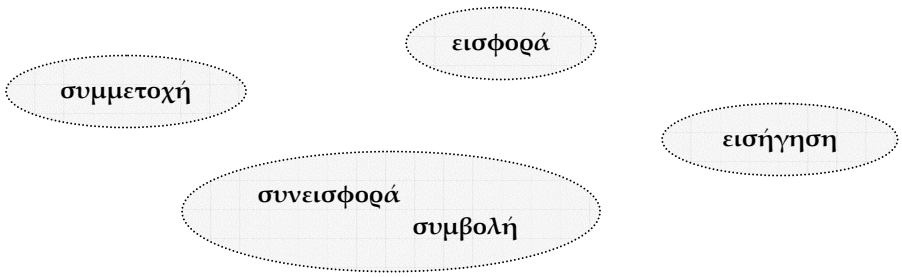


Figure 23. Clusters des EQVs de *contribution*

Le cluster à deux éléments ({*συνεισφορά, συμβολή*}) décrit le sens d'« apport à une œuvre collective ». Le cluster qui contient l'EQV *συμμετοχή* désigne le sens de « participation à une œuvre collective », tandis que le cluster de l'EQV

εισφορά désigne le sens de « cotisation ». Le sens décrit par le cluster de l'EQV *εισήγηση* est le sens de « contribution scientifique » (par ex. un article). Les sens repérés pour *contribution* peuvent être décrits ainsi:

a. *contribution – συμμετοχή*
b. *contribution – εισφορά*
c. *contribution – {συνεισφορά, συμβολή}*
d. *contribution – εισήγηση*

Etant donné que des clusters se chevauchant n'ont pas été créés pour ce mot, la modification de la granularité des sens obtenus n'est donc pas possible.

3.2.4. Clustering sémantique pour le mot *power*

Les EQVs de *power* sont : *δύναμη* (dynami), *αρμοδιότητα* (armodiotita), *ισχύς* (ischys) et *εξουσία* (eksousia). Les clusters constitués à partir de ces EQVs sont décrits dans la figure 24.

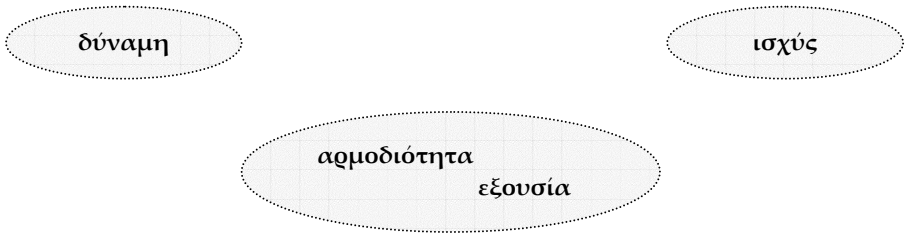


Figure 24. Clusters des EQVs de *power*

Le premier cluster ({*δύναμη*}) décrit le sens de « puissance physique », le cluster à deux éléments {*αρμοδιότητα, εξουσία*} celui de « pouvoir et le cluster {*ισχύς*} celui d'« énergie ». Voici les sens induits pour *power* :

a. *power – δύναμη*
b. *power – ισχύς*
c. *power – {αρμοδιότητα, εξουσία}*

Comme dans le cas de *contribution*, la modification de la granularité des sens n'est pas possible.

3.2.5. Clustering sémantique pour le mot *maintenance*

Maintenance a trois EQVs dans le lexique : *διατήρηση* (diatirisi), *διατροφή* (diatrofi) et *συντήρηση* (syntirisi), qui sont clustérisés de la manière suivante :

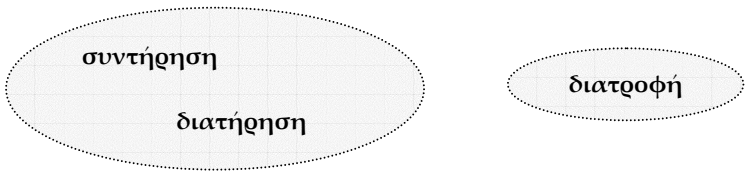


Figure 25. Clusters des EQVs de *maintenace*

Le cluster {*συντήρηση, διατήρηση*} désigne le sens d' « entretien » (ou de « maintien »), tandis que le cluster {*διατροφή*} désigne le sens de « pension alimentaire ». Les sens repérés pour *maintenance* sont donc les suivants :

a. *maintenance* – {*συντήρηση, διατήρηση*}

b. *maintenance* – *διατροφή*

La modification de la granularité de ces sens n'est, là aussi, pas possible.

4. Conclusions sur la méthode d'acquisition de sens

4.1. Points forts de la méthode d'acquisition de sens

La méthode d'acquisition de sens proposée procède par clusterisation des EQVs d'un mot ambigu sur la base de leur similarité sémantique. Le repérage des sens du mot source se fait par projection inter-langue des informations de clustering. Le simple fonctionnement de la méthode met en évidence certains avantages assez clairs.

4.1.1. D'un point de vue opératoire

4.1.1.1. Une méthode non supervisée

Tout d'abord, il s'agit d'une méthode non supervisée, qui n'a donc pas besoin de données sémantiquement pré-étiquetées pour fournir des résultats pertinents. Par conséquent, elle peut être utilisée sur un corpus parallèle, ayant subi certaines étapes de prétraitement automatique. Ces prétraitements consistent en la lemmatisation et l'annotation morphosyntaxique du corpus et en son alignement aux niveaux des phrases et des mots. La lemmatisation et l'annotation morphosyntaxique sont des tâches automatisées pour un grand nombre de langues, pour lesquelles des outils appropriés sont disponibles[35]. Il en va de même pour l'alignement phrastique et lexical ; les outils disponibles sont statistiques, ce qui les rend indépendants de la langue[36]. Ainsi, la méthode peut facilement être appliquée à d'autres paires de langues.

4.1.1.2. Méthode dirigée par les données

En outre, la méthode d'acquisition de sens est dirigée par les données, ce qui permet d'acquérir des sens relatifs aux domaines des textes traités. Cette caractéristique de la méthode empêche l'inclusion de sens non relatifs aux sujets traités dans la base de sens générée, ce qui peut s'avérer profitable lors de l'exploitation de cette ressource pour le traitement. Par exemple, la non considération de tels sens peut « alléger » le processus de WSD ; l'algorithme de WSD visant à identifier le sens véhiculé par une nouvelle instance d'un mot ambigu, n'aura donc pas à éliminer des sens de ce mot non pertinents pour les textes traités. La base de sens peut être mise à jour et enrichie par l'inclusion de nouveaux textes dans le corpus et le relancement du processus d'acquisition de sens.

[35] Ainsi, le TreeTagger (Schmid, 1994), qui permet l'étiquetage morphosyntaxique et la lemmatisation de textes dans un grand nombre de langues, le Brill Tagger (1995), le TnT Tagger (Brants, 2000). Les fichiers des paramètres nécessaires pour le fonctionnement de ces étiqueteurs sont disponibles ou peuvent être générés, pour une langue donnée, à partir de corpus annotés (et lemmatisés) de cette langue.

[36] Ainsi, les aligneurs de phrases et de mots GIZA++ (Och et Ney, 2000) et Uplug (Tiedemann, 2003), et les aligneurs de phrases GMA (Melamed, 1996) et Vanilla (Danielsson et Ridings, 1997).

4.1.1.3. Clustering flou

Du point de vue computationnel, l'algorithme de clustering du programme SEMCLU permet d'effectuer un clustering « flou », où des clusters différents peuvent se chevaucher. Cette possibilité n'est pas offerte par les algorithmes de clustering hiérarchiques ou de partitionnement, qui fournissent des clusters distincts. En revanche, les algorithmes qui permettent un clustering « flou » (par exemple, le *fuzzy c-means*) nécessitent souvent la définition du nombre de clusters obtenus (*c*) dès le début du processus. Cette propriété de la méthode de clustering proposée ici, à savoir la possibilité de recouvrement des clusters générés, a un effet direct sur la description des sens obtenus.

Ainsi, un mérite important de la méthode est de permettre la description et la prise en compte des **relations** éventuellement existantes **entre les sens** du mot ambigu. Si une relation existe entre les sens repérés à l'aide de clusters différents, elle devient alors évidente par le recouvrement des clusters qui décrivent chacun des sens. La possibilité de créer des **clusters se chevauchant** par l'algorithme de clustering permet la prise en compte du **statut** des sens lexicaux, en effectuant une distinction entre les sens **antagonistes**, décrits par des clusters complètement disjoints, et les sens **plus proches** (ou les nuances de sens), décrits par des clusters qui se chevauchent. L'absence de distinction entre sens proches et sens distincts constitue, comme nous l'avons déjà souligné, un reproche souvent adressé aux méthodes classiques d'énumération de sens.

4.1.2. D'un point de vue théorique

4.1.2.1. Hypothèse distributionnelle dans un cadre bilingue

Sur un plan théorique, il s'agit d'une méthode qui sert à estimer la possibilité d'application, dans un contexte de traduction, des hypothèses sous-jacentes à une catégorie de méthodes d'acquisition de sens fortement répandue, à savoir les méthodes distributionnelles. Des méthodes de ce type ont été développées et appliquées essentiellement dans un cadre monolingue. Les travaux qui portent sur l'applicabilité de ces méthodes dans un cadre impliquant

plus d'une langue sont limités, et il en va de même pour des méthodes clairement orientées vers des applications bi- et multi-lingues. La méthode proposée ici permet l'induction de distinctions sémantiques au niveau des mots de la LS, en établissant des correspondances sémantiques inter-langues pertinentes pour la traduction, sur la base de l'hypothèse contextuelle (ou distributionnelle) du sens.

4.1.2.2. Sens de granularité variable

En effet, la méthode proposée permet le repérage de distinctions sémantiques de **granularité variée**. Il est alors possible de considérer des distinctions grossières, comme celles trouvées au niveau des homonymes, aussi bien que des distinctions de granularité fine, concernant des sous-sens ou des nuances sémantiques des mots. Le niveau de granularité des sens retenus peut être sélectionné en fonction des besoins de l'application visée en matière de WSD[37]. Les sens de granularité grossière peuvent être décrits en regroupant des clusters se chevauchant. Mais ce processus implique un certain risque à cause de l'éventuelle ambiguïté des éléments se trouvant à l'intersection des clusters ; il peut néanmoins servir à effacer des distinctions sémantiques erronées, établies en raison de la dispersion des données. Les sens de granularité fine pourraient être repérés, au contraire, par la prise en compte des distinctions présentes entre clusters se chevauchant.

Cette possibilité de variation de la granularité des sens proposés, en fonction des besoins, permet une analyse à **profondeur variable** (Kayser et Coulon, 1989). Dans un tel cadre, la quantité des informations relatives aux mots traités qui sont accessibles peut être modifiée à chaque niveau de profondeur de l'analyse. En ce qui concerne les distinctions sémantiques, la modification de la granularité des clusters de sens permet d'augmenter ou de restreindre le nombre de sens d'un mot source par rapport auxquels une décision doit être prise. En ce qui concerne les EQVs, la quantité des informations disponibles pour un EQV est également

[37] D'après Ide (2006, lors de la communication « Making Senses » à l'Université de Bergen), les besoins de désambiguïsation dans la Recherche d'Information (RI) et dans la Traduction Automatique (TA) pourraient, par exemple, être décrits comme deux extrêmes d'un continuum : la RI nécessite en général une désambiguïsation lexicale « de surface », tandis que la TA peut avoir besoin d'une désambiguïsation plus précise pour générer une traduction qui a l'air plus ou moins naturelle.

modifiée, étant donné que la variation de la granularité des clusters de sens entraîne une augmentation ou une restriction du nombre de mots lui étant sémantiquement liés.

Notre méthode offre précisément cette possibilité de variation de la quantité des informations disponibles. En revanche, un choix doit être opéré relativement au niveau d'analyse adopté au sein d'une application, en fonction des besoins en matière de désambiguïsation (prise en compte des distinctions sémantiques grossières ou fines). Ce niveau ne peut varier de manière automatique.

4.1.2.3. Exploitation de contextes élargis

L'utilisation de contextes « élargis » pour les comparaisons effectuées par le calcul de similarité est considérée comme ayant un effet positif sur les résultats de la méthode. L'utilisation des contextes individuels des instances des mots ambigus, dans une méthode utilisant des cooccurrences de premier ordre, génère des découpages sémantiques très fins. Les distinctions proposées dans ce cas représentent généralement des usages des mots et ne permettent donc pas une généralisation de ces usages en sens. Les contextes élargis, utilisés par notre méthode d'acquisition de sens, sont constitués à partir des segments source et cible correspondants aux EQVs d'un mot ambigu (cf. §1.3). De cette manière les contextes individuels du mot sont regroupés et la possibilité de découvrir des correspondances entre les contextes regroupés augmente, et ce, malgré l'effet omniprésent de la dispersion des données sur le fonctionnement d'une telle méthode (basée sur des cooccurrences de premier ordre).

4.1.2.4. Prise en compte d'informations paradigmatiques

Pour conclure, un autre point fort de la méthode proposée est de permettre l'enrichissement des correspondances de traduction établies entre les mots de deux langues par des **informations paradigmatiques**. Ces informations concernent les relations de similarité sémantique entre les EQVs clustérisés : ces

EQVs sont souvent des *quasi-synonymes*[38] et leurs relations peuvent être exploitées lors du processus de prédiction de traduction. En fonction de leur similarité, les EQVs d'un cluster peuvent être plus ou moins substituables en tant que traductions du sens du mot source décrit par le cluster en question. En effet, même lors d'une tâche de traduction manuelle, le traducteur est fréquemment confronté à des possibilités multiples, entre lesquelles le choix ne s'avère pas toujours évident. Ainsi, la création de correspondances bi-univoques entre sens du mot source et EQVs (du type '*un sens ◎ un EQV*') paraîtrait peu naturelle et assez réductrice.

La manière dont les correspondances sémantiques de ce type peuvent être exploitées, de manière profitable, lors de la prédiction de traduction pour de nouvelles instances des mots ambigus sera analysée dans le chapitre 8. Nous montrerons que la considération des informations paradigmatiques mises en évidence a également un effet bénéfique sur les résultats de l'évaluation du processus de prédiction de traduction.

4.2. Points faibles de la méthode d'acquisition de sens

4.2.1. D'un point de vue opératoire

4.2.1.1. Sensibilité à la dispersion des données

L'inconvénient principal de la méthode d'acquisition de sens proposée est sa vulnérabilité aux effets de la dispersion des données.

Les conclusions sur la sémantique des mots dépendent en effet fortement des informations distributionnelles trouvées dans le corpus. La méthode est complètement dirigée par les données et exploite des **informations de surface**, relatives aux **cooccurrences de premier ordre** des mots. Ainsi il arrive que des distinctions sémantiques erronées soient établies, en raison du manque d'une forte similarité au niveau des mots entre les contextes correspondants aux EQVS. Des informations plus abstraites, qui permettraient d'établir des relations de similarité entre contextes sémantiquement similaires mais composés de mots

[38] Etant donné la rareté du phénomène de synonymie absolue dans les langues naturelles.

différents (par ex. de cooccurrences d'ordre plus élevé), ne sont pas ici exploitées ; l'exploitation d'informations de ce type pourrait ainsi constituer une piste de recherche future.

4.2.1.2. Sensibilité au bruit présent dans les résultats de l'alignement

La performance de la méthode d'acquisition de sens dépend fortement de la qualité des résultats de l'alignement des mots. Si ces résultats contiennent du bruit, c'est-à-dire des EQVs erronés (ce qui correspond à une faible précision et, souvent, à un haut rappel), ces EQVs seront pris en compte par le processus de clustering et le bruit sera également repéré dans les clusters de sens récupérés en sortie.

Au contraire, une haute précision et un bas rappel dans les résultats de l'alignement privilégient la qualité des informations traductionnelles fournies au détriment de leur quantité, qui parfois n'est pas suffisante. Dans ce cas, il se peut que très peu d'EQVs possibles d'un mot dans le corpus soient repérés. Comme nous l'avons déjà souligné, la performance de la méthode est d'autant meilleure que le nombre d'EQVs des mots ambigus est élevé ; avec un nombre d'EQVs très petit (deux ou trois), la méthode ne parvient pas à donner des résultats pertinents. Par conséquent, les résultats de l'alignement automatique doivent être caractérisés à la fois par une **haute précision** et un **haut rappel**. L'existence de ce problème justifie à lui seul l'utilisation de données manuellement repérées.

Fournir de bons résultats concernant ces deux aspects n'est pourtant pas évident pour un aligneur automatique. L'objet de notre travail n'étant pas l'amélioration de l'alignement lexical automatique, nous avons mené des expériences sur des données de traduction manuellement repérées, afin d'estimer les possibilités de la méthode dans le cas de données automatiquement générées de haute qualité. Les résultats sur les données manuellement repérées s'étant révélés très pertinents, nous croyons que la performance de la méthode serait satisfaisante si des données automatiquement générées de haute qualité étaient disponibles.

4.2.2. D'un point de vue théorique

4.2.2.1. Analyse non exhaustive de la sémantique des EQVs

Du point de vue de l'analyse sémantique, la région sémantique couverte par le mot source est bien décrite par sa mise en correspondance avec des mots pouvant véhiculer les mêmes sens dans la LC. Si cette analyse permet d'avoir une image claire de la sémantique du mot source, ce n'est pas le cas pour les EQVs.

En effet, l'étude de la sémantique des EQVs est limitée aux sens relatifs à la sémantique du mot ambigu. Il se peut pourtant que les EQVs soient également des mots ambigus dans la LC et que seule une partie de leur *potentiel sémantique* se manifeste au sein du sous-corpus d'un mot source donné. Ainsi, l'analyse sémantique effectuée ne concerne pas la partie de leur région sémantique qui n'est pas liée à la sémantique du mot source et qui ne se manifeste pas dans son sous-corpus. Il n'est donc pas possible de savoir quelle partie du potentiel sémantique de l'EQV occupe un sens donné.

Néanmoins, la non prise en compte de l'ambiguïté au sein de la LC n'a pas d'effet négatif dans une étude traductionnelle directionnelle, telle que celle que nous menons ici ; les correspondances de traduction pour les mots ambigus source étudiés sont décrites de manière à permettre leur modélisation et leur utilisation efficace dans un cadre automatique. L'absence d'une analyse de la polysémie des mots de la LC ne constituerait un problème que dans le cas d'une inversion de la direction de traduction.

Pour modéliser toute la gamme des solutions de traduction pour les mots ambigus de la LC – qui deviendraient, à leur tour, des mots source –, il faudrait analyser leurs autres sens à l'aide d'EQVs repérés dans le reste du corpus. Ainsi, des processus de construction de sous-corpus et de filtrage devraient avoir lieu pour les mots de la LC (cf. §1.3.10 et §1.3.11, chapitre 4). Hormis l'intérêt pratique de la question, créer des correspondances sémantiques entre l'ensemble des régions occupées par les mots de deux langues et étudier la gamme complète de traductions possibles pour les mots ambigus serait intéressant également au niveau théorique.

4.2.2.2. Absence de spécification des relations entre EQVs clustérisés

Un autre inconvénient de la méthode résulte du fait que le clustering obtenu ne permet pas de spécifier le type des relations entre les EQVs clustérisés. Le calcul distributionnel permet le repérage de relations de similarité sémantique entre les EQVs, mais ces relations peuvent être de types variés. Nous considérons qu'il serait intéressant et utile d'analyser la nature de ces relations ; une telle analyse permettrait d'aboutir à des conclusions concernant la substituabilité des EQVs similaires en contexte.

Afin d'analyser la nature des relations (d'une partie) des EQVs clustérisés, nous avons recouru à une autre méthode d'analyse sémantique, la méthode des Miroirs Sémantiques (Dyvik, 2003, 2005), qui exploite des informations de traduction pour la construction d'un thésaurus sémantique[39].

4.2.2.3. Risques lors de la construction de sens de granularité grossière

Enfin, un autre inconvénient que nous avons observé concerne les risques de construire des sens à granularité grossière. Comme nous l'avons montré, des sens plus **grossiers** peuvent être obtenus en **fusionnant**, lors d'une étape de post-traitement, les **clusters se chevauchant** issus des résultats de SEMCLU, fusionnement qui peut avoir lieu même s'il manque certaines relations entre les clusters.

Le risque évident, dans une telle approche, est de regrouper des sens distincts, en raison de l'ambiguïté des EQVs situés à l'intersection des clusters. Néanmoins, si ce regroupement n'est pas effectué et si chaque cluster d'un ensemble de clusters se chevauchant décrit un sens, les sens proposés sont parfois de granularité trop fine ; ils correspondent plus à des nuances sémantiques qu'à de véritables sens des mots. La raison principale de ce phénomène est la dispersion des données, problème omniprésent dans les travaux d'acquisition de sens sur la base de la distribution lexicale, qui s'intensifie dans le cas d'utilisation d'informations de surface (cooccurrences de premier ordre).

[39] L'application de cette méthode à nos données et les résultats obtenus seront présentés plus loin.

CONCLUSION

Dans ce chapitre, nous avons présenté la méthode d'acquisition de sens développée dans ce travail, en commençant par les hypothèses théoriques sous-jacentes à la méthode, qui permettent l'extension de l'hypothèse distributionnelle du sens dans un cadre de traduction. Ensuite, nous avons présenté le fonctionnement du programme SEMCLU, qui effectue un clustering sémantique des EQVs des mots ambigus. La projection des informations des clusters générés sur les mots ambigus source permet le repérage des sens véhiculés par ces mots au sein du corpus d'apprentissage.

Notre méthode d'acquisition de sens étant une méthode d'apprentissage non supervisé complètement dépendante des données, l'analyse sémantique obtenue dépend fortement des informations contextuelles et traductionnelles trouvées dans le corpus. Les résultats obtenus par cette méthode sur des données de traduction manuellement et automatiquement acquises ont été illustrés à l'aide d'un échantillon de mots. Les résultats obtenus pour l'ensemble des données analysées sont présentés en Annexes D3 et D4.

Les résultats du processus d'acquisition de sens peuvent être exploités pour la désambiguïsation lexicale de nouvelles instances des mots ambigus source, ainsi que pour la prédiction de la traduction la plus adéquate de ces nouvelles instances, lors d'une tâche de sélection lexicale. Les méthodes permettant d'effectuer ces types de traitement, développées et implémentées dans le cadre de ce travail, seront présentées dans le chapitre 8.

CHAPITRE 7

SIMILARITE SEMANTIQUE

INTRODUCTION

Jusqu'ici nous avons fait référence (à plusieurs reprises) aux notions de **similarité lexicale** et **sémantique** : des relations de similarité sémantique sont proposées par le calcul de similarité distributionnelle, relations qui servent par la suite au clustering des équivalents (EQVs) des mots ambigus. Pourtant, la nature des relations pouvant être décrites comme des relations de similarité sémantique n'a pas été analysée. Sous quels aspects les mots peuvent-ils être similaires entre eux et quelles sont les conséquences de la nature de leur similarité sur leur traitement ? En outre, est-ce que la force des relations repérées entre les mots peut avoir un impact sur la manière dont ils sont traités ?

La question du traitement de mots sémantiquement proches est fortement liée à celle de leur **substituabilité** (ou permutabilité). Le principe de permutation

des mots similaires en contexte est souvent associé à la constitution de classes sémantiques, comme celles fournies par les méthodes distributionnelles. Néanmoins, l'utilité de ces classes et de ce principe pour émettre des jugements sur la permutabilité des mots en contexte demeure limitée, sans informations sur ce qui définit la spécificité des sens des mots regroupés, au-delà de leur sens commun (Rossignol et Sébillot, 2006).

Les notions de similarité sémantique et de substituabilité continueront à constituer des notions centrales au sein des méthodes de désambiguïsation et de sélection lexicale, qui seront présentées dans le chapitre 8. Avant d'y parvenir, nous considérons comme important de clarifier davantage ces notions d'un point de vue théorique, et d'analyser la manière dont elles peuvent être prises en compte dans les systèmes de TAL.

1. Relations sémantiques entre unités lexicales

1.1. Similarité sémantique : une notion vague

La **similarité lexicale** ou **sémantique** est une notion très vague qui peut être utilisée pour décrire des relations diverses entre les mots, aussi bien que des types variés de relations lexicales. Sparck Jones (1986 : 42-47) donne une liste (non exhaustive) des relations proposées ou illustrées dans des classifications sémantiques : les rapports associatifs de Saussure ou l'analogie de concepts ; la synonymie, comme elle est conçue dans les dictionnaires de synonymes et illustrée par les groupes de mots au sein des descripteurs des thésaurus ; l'association, notion sous-jacente au champ associatif de Bally (1940) ; les relations d'appartenance à une classe (membre, espèce, objet) ; l'hyponymie (Lyons, 1968 : 453-455), déterminée en termes d'implication et alternativement décrite comme inclusion ; l'antonymie ; l'incompatibilité (exclusion) (Lyons, *ibid*.) ; la conséquence ; la collocation (Firth, 1957b) ; l'analogie et, finalement, la relation entre un mot très général, ou classificateur, et d'autres mots plus spécifiques. Nous pourrions ajouter à cette liste les relations de méronymie et de quasi-synonymie.

Sovran (1992) souligne également le vague de la notion de similarité lexicale mais il préfère parler d'un **cluster de notions** plutôt que d'un concept simple. Pour lui, il existe une unité sous-jacente aux différents sous-types de similarité qui justifie leur rassemblement dans un concept général. Malgré le vague de la notion, le besoin d'une définition opérationnelle et précise des relations sémantiques apparaît dans un cadre pratique, définition permettant tant la caractérisation des mots par rapport à leur relation que le repérage des mots liés par une relation précise[1].

La similarité sémantique des mots peut être définie hors contexte ou en fonction du contexte dans lequel ils sont utilisés. La première approche se retrouve dans la discussion classique sur la question. Dans la théorie des champs sémantiques (Trier, 1934), par exemple, qui se situe dans le courant structuraliste, les mots sémantiquement apparentés sont conçus comme formant un champ. Ce champ sémantique peut être considéré comme une mosaïque, où la taille de chaque pièce est déterminée par celle de ses voisins ; ainsi, l'extension du sens d'un mot à l'intérieur d'un champ est considérée comme impliquant une restriction correspondante du sens de ses voisins. Une telle conception stricte des sens lexicaux qui implique leur exclusion mutuelle n'est pourtant pas justifiée ; les sens des mots peuvent effectivement présenter des recouvrements ou avoir des frontières vagues (Sparck Jones, 1986 : 39). La conception d'un champ comme mosaïque présuppose aussi que les mots ont un sens constant à chaque moment, indépendant des contextes dans lesquels ils sont utilisés.

Néanmoins, la région sémantique des mots n'est pas toujours clairement définissable et l'énumération de leurs différents sens est souvent difficile, justement en raison des recouvrements qui peuvent exister entre les sens. L'**usage** des mots constitue une notion centrale des travaux en sémantique lexicale, ce qui rend l'étude dépendante du contexte. Lorsque le sens est conçu de cette manière, la similarité des mots ne peut découler que de l'étude des usages contextualisés des mots. L'identification des relations lexicales devient ainsi dépendante du contexte.

[1] Il ne suffit pas de pouvoir caractériser une relation entre deux éléments *A* et *B* ; il faut aussi connaître la procédure permettant de repérer *B* étant donné *A*.

La question de la similarité sémantique n'est pourtant pas une question d'ordre purement linguistique. Nous la retrouvons dans les domaines de la psycholinguistique, de la psychologie et de la linguistique cognitives.

1.2. Similarité sémantique dans un cadre cognitif

1.2.1. Approche en termes de distance mentale

Dans un cadre cognitif, la similarité sémantique des mots décrit souvent leur relation d'**association**, à savoir que la présence de l'un induit l'activation mentale de l'autre (Lemaire et Denhière, 2006). Les jugements de similarité sémantique concernent ainsi la force d'association des mots et dépendent de leur fréquence de cooccurrence ; leur similarité est considérée d'autant plus grande qu'ils apparaissent plus souvent ensemble.

Dans ce champ de recherche, la similarité sémantique des mots fait souvent référence à la proximité psychologique des représentations mentales correspondantes et est liée à l'étude du **traitement lexical** (Tabossi, 1989). Une approche très importante dans ce cadre est l'**approche en termes de distance mentale**, qui exploite la notion d'espace mental (Shepard, 1962) : les concepts sont représentés en tant que points dans l'espace et leur similarité est une fonction de leur distance, calculée à l'aide de techniques d'**échelonnement multidimensionnel** (Arnold, 1971). Les concepts correspondants à des points qui se trouvent à proximité sont considérés comme plus similaires, d'un point de vue psychologique, que les points distants. Le lexique mental peut également être regardé comme un réseau, dont les entrées (qui correspondent aux mots) sont liées par des relations variées. Là aussi, les mots apparentés se trouvent plus près que ceux qui n'entretiennent pas de relations fortes.

Le traitement lexical au sein de ces modèles est considéré comme influencé par le contexte des mots : lorsqu'un mot est reconnu, l'entrée correspondante dans le réseau est activée et l'activation se propage automatiquement aux entrées à proximité (Collins et Loftus, 1975). Ce phénomène, souvent appelé **amorçage** (*lexical priming*), se manifeste dans le temps plus court requis pour identifier le sens d'un mot, lorsque ce mot suit, dans le texte, un mot sémantiquement

apparenté (mot amorce) (Meyer *et al*., 1975 ; Seidenberg *et al*., 1984 ; Le Ny, 1989). Cette possibilité de prise en compte du temps nécessaire pour l'identification du sens rend les modèles d'activation compatibles avec des faits expérimentaux qui impliquent des mesures de temps de réponse de diverses sortes (Le Ny, *ibid*.).

1.2.2. Approche de traits ou de contraste

Une autre approche importante dans le domaine de la psychologie cognitive est l'**approche en termes de traits** ou **de contraste**. L'hypothèse centrale de cette approche est que les concepts peuvent être représentés par des listes de traits, décrivant des propriétés des objets correspondants. Une comparaison de similarité entre concepts consiste donc en une comparaison de leurs listes de traits[2]. Leur similarité dépend du degré de recouvrement de leurs ensembles de traits (Tversky, 1977) et peut être conçue comme une fonction croissante de leurs traits communs et comme une fonction décroissante de leurs traits distinctifs[3].

Un corollaire de ce principe est qu'un concept peut être similaire à un autre selon un aspect particulier, désigné par les traits qui leur sont communs. Cette hypothèse se rapproche de la notion d'**air de famille** proposée par Wittgenstein (1953) ; selon cette notion, des entités différentes peuvent être liées par une **ressemblance de famille**, représentée par le recouvrement d'un ensemble de leurs propriétés, aucune propriété n'étant commune à toutes. Cette conception de la similarité révèle son caractère **non transitif** : si deux mots A et B sont similaires à C, il n'en découle pas pour autant que A et B sont similaires entre eux, dans la mesure où leur similarité à C peut impliquer des aspects différents. Cela ne signifie pas, néanmoins, qu'une telle relation ne puisse exister entre A et B, mais elle doit alors être établie indépendamment de C.

Une telle conception de la notion de similarité constitue pourtant un écueil possible pour la définition de la notion de similarité, qui risque ainsi d'être privée

[2] Si l'on considère que la distance sémantique peut se définir comme une fonction des traits composant les concepts, l'approche en termes de comparaison de traits est alors équivalente à celle de propagation d'activation (Tversky, 1977 ; Cornuéjols, 2003 : 63).

[3] Cette approche a été adoptée au sein de notre travail : l'estimation de la similarité sémantique des EQVs des mots ambigus est faite sur la base de leurs ensembles de traits contextuels (source ou cible) retenus.

de sens étant donné que toutes les entités peuvent être considérées comme ayant une relation quelconque. Des moyens de restriction de la notion de similarité ont été proposés. Ainsi le recours à la notion de **pertinence**, qui impose la considération uniquement des traits saillants qui caractérisent les objets comparés. Selon Tversky (1977), la saillance ou pertinence des traits est déterminée par leur intensité et leur diagnosticité, c'est-à-dire le degré auquel ils sont significatifs pour un certain jugement de similarité. Sperber et Wilson (1989 : 190, 197-199) soulignent aussi que la pertinence est une question de degré, qui ne devrait pas être formulée comme un **concept classificatoire** (en des termes de conditions nécessaires et suffisantes) ou **quantitatif**, mais en tant que **concept comparatif**. La notion comparative de pertinence est caractérisée en termes d'effet et d'effort (Sperber et Wilson 1989 : 230). Un stimulus est un phénomène mis en place afin de produire des effets cognitifs particuliers ; dans une définition comparative de sa pertinence, le phénomène est considéré d'autant plus pertinent que les effets contextuels qu'il produit, lorsque traité de manière optimale, sont importants, ou que l'effort nécessaire pour le traiter de manière optimale est faible.

Une autre restriction possible de la notion de similarité peut venir de la conception même de cette notion. L'approche basée sur les traits et l'approche en termes de distance mentale traitent la similarité comme un prédicat à deux places (*S(a,b)* : *a* est similaire à *b*). Cette conception de la similarité a été pourtant critiquée comme ne restreignant pas suffisamment les manières dont les objets peuvent être similaires ou différents. Pour Goodman (1970, 1972), la « similarité d'un élément *A* à un élément *B* » est une notion privée de sens, sauf si les aspects dans lesquels *A* est similaire à *B* sont décrits. La similarité est donc, pour Goodman, un prédicat à trois places (« l'élément *A* est similaire à *B* en ce qui concerne *C* »). Pour Medin et Goldstone (1995 : 106) même cela n'est pas suffisant ; quant à eux, ils considèrent la similarité comme un prédicat à plusieurs places, ce qui permet d'enrichir les jugements de similarité par des informations supplémentaires (sur le processus de comparaison, un standard, la perspective). Goldstone *et al*. (1997) soulignent même le rôle du contexte du jugement de similarité sur le jugement lui-même.

La **théorie du prototype** (Rosch, 1978 ; Kleiber, 1990 ; 1999 : 59-61) a essayé aussi de restreindre la notion de similarité. Au sein de cette théorie, les traits des éléments comparés sont conçus comme présents ou absents à un certain degré. La similarité des éléments est ensuite évaluée par considération de leur proximité relative à un prototype – qui sert en tant que ***tertium comparationis*** – en termes de partage de traits[4].

Dans le paragraphe suivant, nous présenterons certaines manières de prendre en compte la similarité lexicale dans un cadre automatique.

2. Considération de la similarité sémantique dans un cadre automatique

2.1. Traitement basé sur la *similarité*

2.1.1. Méthodes basées sur les connaissances

2.1.1.1. Approche de contenu informationnel

Les méthodes d'estimation de la similarité sémantique basées sur les connaissances exploitent des ressources préétablies. Le plus souvent, ces ressources sont des hiérarchies lexicales, comme le réseau lexico-sémantique **WordNet**. Au sein d'un tel réseau, les mots (concepts) sont représentés par des nœuds, tandis que les relations entre les concepts sont décrites par des arêtes liant les nœuds entre eux. L'estimation de la similarité peut être basée sur les nœuds qui représentent les concepts ; une telle approche est appelée « approche de **contenu informationnel** ».

Chaque nœud du réseau décrit un seul concept, qui consiste en une quantité précise d'informations. L'arête liant deux nœuds représente une association directe entre les concepts correspondants (Resnik, 1995 ; 1999). La similarité de

[4] L'application de la théorie du prototype dans le cas de la polysémie concerne la possibilité de regarder une unité lexicale polysémique comme une catégorie prototypique. L'appartenance à une telle catégorie, c'est-à-dire l'apparentement des sens différents, est fondée sur une relation de ressemblance de famille.

deux concepts, dans une approche de contenu informationnel, est ainsi considérée comme proportionnelle à leurs informations communes. Dans un espace conceptuel hiérarchique, le « porteur » des informations communes de deux concepts est le concept qui les subsume au sein de la hiérarchie (par exemple, leur hyperonyme).

2.1.1.2. Approche en termes de distance conceptuelle

L'estimation de la similarité sémantique des mots à l'aide d'un réseau lexico-sémantique peut se baser aussi sur la distance des arêtes liant les nœuds entre eux. Cette approche est appelée approche de **distance conceptuelle**. Dans une telle approche, la similarité sémantique de deux concepts est mesurée sur la base de la distance entre leurs nœuds au sein de la hiérarchie, calculée le plus souvent en nombre d'arêtes liant les nœuds[5] (Leacock et Chodorow, 1998 ; Lee *et al.*, 1989 ; Rada *et al.*, 1989 ; Wu et Palmer, 1994 ; Agirre et Rigau, 1995). L'approche de distance est considérée comme une manière naturelle, directe et intuitive d'estimation de la similarité sémantique dans une taxonomie. D'autres paramètres pouvant être pris en compte dans cette approche, à part la longueur du chemin liant deux nœuds, sont :

a. la **profondeur** des concepts dans la hiérarchie : les concepts qui se trouvent dans une partie plus profonde de la hiérarchie sont considérés comme plus proches
b. la **densité** des concepts dans la hiérarchie : ceux qui sont situés dans une partie dense de la hiérarchie sont relativement plus proches que ceux qui se trouvent dans une région plus clairsemée.

Un inconvénient de cette méthode est sa dépendance vis-à-vis de la hiérarchie du réseau, qui est prédéfinie sur la base de jugements subjectifs[6]. Les

[5] Ces méthodes ont des points communs avec le modèle de distance mentale, développé dans le cadre de la psychologie cognitive (cf. 1.2.1.).
[6] L'approche de contenu informationnel nécessite moins d'informations sur la structure détaillée de la taxonomie tout en lui restant, elle aussi, dépendante. Néanmoins, le faible degré de prise en compte des informations sur la structure est parfois considéré comme une des faiblesses de la méthode.

ressources manuellement créées de ce type sont en général caractérisées par une couverture limitée, en raison du temps requis pour leur construction. Ainsi, la plupart des ressources disponibles concernent des domaines bien précis (à l'exception de WordNet), ce qui rend leur exploitation difficile pour le traitement de la langue générale. Des efforts de dérivation automatique de telles ressources à partir de dictionnaires existent (Chodorow *et al*., 1985), mais l'utilisation de telles méthodes ne constitue pas une pratique commune. La nécessité de représentation du domaine en termes de réseau est considérée comme une faiblesse importante des méthodes basées sur la distance (Lin, 1998a), qui ne peuvent pas s'appliquer dans le cas de non disponibilité d'une telle représentation.

2.1.1.3. Mesures de type Lesk

Une autre manière d'estimation de la similarité sémantique des mots implique le repérage de recouvrements (en termes de mots partagés) entre leurs définitions dictionnairiques. Il s'agit de mesures de type Lesk (1986)[7] : ces mesures considèrent les sens des mots comme d'autant plus liés que le nombre de mots partagés par leurs définitions est élevé.

Banerjee et Pedersen (2003) utilisent cette mesure de similarité comme complément d'une autre, basée sur la hiérarchie de WordNet. La combinaison des informations trouvées dans les définitions des concepts fournies dans WordNet et d'informations hiérarchiques permet de construire des **définitions étendues** des sens. Cette extension des définitions des concepts consiste en leur enrichissement par des informations trouvées dans les définitions d'autres concepts, qui leurs sont liés dans la hiérarchie (par ex. leurs hyperonymes). Le recouvrement des définitions dans ces cas est considéré refléter des relations conceptuelles implicites, manquantes du réseau, comme des relations entre les hypéronymes des synsets initialement comparés.

[7] La méthode de désambiguïsation lexicale de Lesk est présentée en détail dans le paragraphe 2.1.2, chapitre 4. Son hypothèse de base est que les mots qui apparaissent dans la même phrase sont utilisés dans des sens proches et que le degré auquel les sens lexicaux sont liés peut être identifié par le nombre de recouvrements dans leurs définitions dictionnairiques.

La mesure de similarité de Banerjee et Pedersen permet un ensemble étendu de comparaisons, qui concernent les recouvrements entre les définitions des synsets hyperonymes, hyponymes et méronymes des synsets dont la similarité doit être estimée. En combinant les avantages des recouvrements des définitions et de la structure d'une hiérarchie conceptuelle, elle crée une notion étendue de la similarité entre synsets. Cette mesure a été reprise dans Patwardan *et al*. (2003) et Pedersen *et al*. (2004) et utilisée dans des buts d'évaluation dans Mihalcea *et al*. (2006) et Sinha et Mihalcea (2007).

2.1.2. Méthodes basées sur les données

Les méthodes d'estimation de la similarité sémantique des mots basées sur les données utilisent des mesures distributionnelles. Les informations utilisées dans ce cas sont extraites de corpus et il n'y a donc pas besoin de ressources préexistantes. L'hypothèse sous-jacente à ces approches est l'hypothèse distributionnelle de la similarité sémantique (Charles et Miller, 1998), d'après laquelle des mots sémantiquement similaires présentent des comportements distributionnels similaires.

Pour effectuer un calcul de similarité sur la base d'informations distributionnelles, les cooccurrents d'un mot dans les textes sont considérés comme ses traits et la similarité de deux mots correspond à la similarité de leurs ensembles de traits. Les traits d'un mot peuvent être représentés sous la forme de vecteur et l'estimation de la similarité de deux mots coïncide avec le calcul de la distance de leurs vecteurs au sein d'un espace multidimensionnel. Dans un tel cadre, les mots qui apparaissent dans des contextes similaires ont des vecteurs similaires, reflétant la grande corrélation de leurs usages.

Une méthode d'estimation de similarité basée sur des vecteurs est l'Analyse Sémantique Latente (Latent Semantic Analysis (LSA)) (Landauer et Dumais, 1997 ; Landauer *et al*., 1998 ; Deerwester *et al*., 1990) ; dans cette méthode, un mot est représenté par un vecteur de traits et la similarité de deux mots est estimée à l'aide du cosinus de l'angle que forment leurs vecteurs. Néanmoins, chaque mot

étant représentée par un seul vecteur, la LSA ne permet pas de représenter la polysémie des mots de manière directe (Pedersen, 2007)[8].

Church et Hanks (1990) et Dagan *et al*. (1993) exploitent des informations de cooccurrence et des métriques basées sur l'information mutuelle[9] pour l'estimation de la similarité. Pantel et Lin (2002) construisent des vecteurs basés sur l'information mutuelle pour les mots comparés et calculent leur similarité à l'aide du cosinus des vecteurs associés. Contrairement à la LSA, l'algorithme de Pantel et Lin ('Clustering by Committee') découvre des synonymes des différents sens d'un mot.

Notre travail est aussi basé sur les données et l'estimation de similarité sémantique des mots se fait à l'aide d'un calcul de similarité distributionnelle (cf. §2.3, chapitre 6). La similarité des EQVs d'un mot ambigu est induite par leur similarité distributionnelle par rapport aux contextes source. Comme nous l'avons déjà souligné, la méthode utilisée permet l'estimation de la force des relations repérées mais pas l'analyse de leur nature, ce qui constitue un inconvénient important des méthodes distributionnelles[10].

Rubenstein et Goodenough (1965) soutiennent la corrélation positive entre le degré de similarité sémantique (synonymie) des mots et le degré de similarité de leurs contextes, mais considèrent la similarité contextuelle comme un critère fiable seulement pour le repérage de relations sémantiques très fortes. En revanche, d'après Charles et Miller (*ibid*.) et Justeson et Katz (1991), la nature des relations induites sur la base de la similarité distributionnelle est variable, pouvant aller de cas de synonymie et de quasi-synonymie jusqu'à des cas

[8] Cette incapacité de représenter la polysémie de manière directe caractérise également un autre modèle de calcul de la similarité sémantique entre des mots, le 'Hyperspace Analogue to Language' (HAL) (Burgess et Lund, 1997).

[9] Etant donné deux mots x et y, l'information mutuelle compare la probabilité de trouver les x et y ensemble (probabilité jointe) avec les probabilités que x et y soient indépendants (Fano, 1961). Les mots sont indépendants si la réalisation de l'un n'apporte aucune information sur la réalisation de l'autre. L'information mutuelle est nulle si les mots sont indépendants et accroît lorsque la dépendance augmente.

[10] Nous considérons que l'identification de la nature des relations entre mots sémantiquement similaires pourrait contribuer à une meilleure estimation de leur substituabilité en contexte. Etant donné que la méthode distributionnelle ne permet pas une telle analyse, nous avons essayé d'explorer cette question en complémentant nos résultats avec ceux d'une autre méthode d'analyse sémantique, qui se sert d'informations de nature différente : il s'agit de la méthode des Miroirs Sémantiques (Dyvik, 2003 ; 2005), qui utilise seulement des informations traductionnelles ne prenant pas en compte le contexte des mots. L'application de cette méthode à nos données et les résultats obtenus seront présentés dans le chapitre 7.

d'antonymie. Selon Resnik (1995), cet inconvénient des méthodes distributionnelles, à savoir le repérage de relations dont la nature est difficile à identifier, est accentué par la difficulté d'interprétation des métriques de similarité utilisées ; ces métriques permettent d'obtenir des classes de mots caractérisées comme sémantiques, mais souvent les similarités qu'elles captent sont syntaxiques, pragmatiques ou stylistiques.

Dans certaines méthodes, les informations contextuelles exploitées ne concernent pas la simple cooccurrence des mots mais les résultats d'une analyse syntaxique (Hindle, 1990 ; Grefenstette, 1992 ; Pereira *et al*., 1993 ; Lin, 1998a,b ; Gasperin *et al*., 2001 ; van der Plas et Bouma, 2004). Les traits des mots sont dans ces cas les contextes syntaxiques dans lesquels ils apparaissent. Lin (*ibid*), par exemple, utilise des triplets de dépendance, dont chacun concerne deux mots et la relation syntaxique qui les lie dans le corpus, et chez Hindle (1990) l'estimation de la similarité sémantique est faite par référence aux structures de type 'prédicat-arguments' dans lesquelles les mots apparaissent[11].

2.1.3. Méthodes hybrides

Des méthodes hybrides d'estimation de la similarité sémantique existent aussi, qui combinent des informations taxonomiques basées sur les connaissances (trouvées dans des ressources prédéfinies) et des informations obtenues à partir de corpus à l'aide de méthodes statistiques (Jiang et Conrath, 1997 ; Resnik, 1995 ; Lin, 1998a). La méthode de Jiang et Conrath, par exemple, combine des informations relatives à la structure d'une taxonomie lexicale avec des informations statistiques, ce qui apporte une amélioration des résultats par rapport à ceux obtenus par les méthodes de contenu informationnel ou basées sur la distance. En effet, l'évidence dérivée d'une analyse distributionnelle sur

[11] Hindle caractérise la classification qui résulte de cette méthode comme « quasi-sémantique » et doute de son utilité directe sans une intervention humaine considérable. Il souligne aussi le besoin d'une méthode de discrimination des sens, parce que la méthode n'arrive pas à distinguer entre les sens différents des mots similaires.

corpus sert à quantifier la distance sémantique entre les nœuds au sein d'un espace sémantique construit à l'aide de la taxonomie[12].

2.2. Traitement basé sur la *différence*

2.2.1. Pourquoi la *différence* est-elle importante ?

Les relations entre les mots estimés sémantiquement similaires à l'aide des méthodes basées sur leurs propriétés communes (ou leurs traits partagés) sont souvent des relations de **quasi-synonymie**, étant donné la rareté du phénomène de synonymie dans les langues naturelles. Néanmoins, dans une étude de la quasi-synonymie, ce qui importe le plus n'est pas les similarités entre les mots comparés mais leurs **différences**, c'est-à-dire leurs traits distinctifs (DiMarco *et al.*, 1993 ; Edmonds, 1999). Cet aspect est également d'une importance primordiale au niveau des applications qui impliquent des tâches de sélection lexicale ; ce sont précisément les différences qui permettent la sélection entre mots sémantiquement similaires. Pourtant, l'étude de ces différences n'est pas évidente, en raison du manque de descriptions sémantiques de granularité d'une telle finesse au sein des ressources lexicales, et de l'imprécision des mesures utilisées (Edmonds et Hirst, 2002).

2.2.2. Approches de la *différence*

Edmonds et Hirst (*ibid*.) proposent un modèle permettant l'estimation de la similarité entre quasi-synonymes qui permet la mise en valeur de leurs différences. Plus concrètement, le sens d'un mot en contexte émerge, dans ce modèle, de la combinaison d'une **dénotation de base** lui étant inhérente et indépendante du contexte, et d'un ensemble de **différences explicites** entre lui et ses quasi-synonymes (qui forment un cluster). La représentation des nuances

[12] Cette métrique a été celle qui a donné les meilleurs résultats dans les expériences de Budanitsky et Hirst (2001), qui concernent la comparaison de cinq métriques de similarité sémantiques exploitant WordNet. Les autres métriques qui ont été testées sont celles proposées par Hirst et St-Onge (1998), Leacock et Chodorow (1998), Resnik (1995), Lin (1998a).

sémantiques des mots dans des termes positifs, absolus et indépendants du contexte étant difficile et parfois même impossible, Edmonds et Hirst proposent de les représenter en tant que différences entre quasi-synonymes, dans le sens saussurien.

Les différences de granularité fine entre deux quasi-synonymes sont donc implicitement encodées dans des ensembles de distinctions qui leur correspondent, à un niveau **sous-conceptuel/stylistico-sémantique**. Ce niveau se situe entre un niveau **conceptuel-sémantique**, où le sens dénotationnel partagé par les quasi-synonymes est représenté à l'aide d'une ontologie, et un niveau de représentation **syntactico-sémantique**, qui contient des informations syntaxiques et de collocation, concernant les combinaisons possibles des mots[13]. Par conséquent, le sens lexical n'est pas représenté de manière explicite dans le lexique, mais créé ou généré lors de son usage. La comparaison des ensembles de distinctions caractérisant les quasi-synonymes est pourtant complexe, à cause du très grand nombre de dimensions sur lesquelles les mots peuvent différer. En outre, il est possible qu'une correspondance entre les distinctions de quasi-synonymes différents n'existe pas, ce qui empêche d'avoir une base pour la comparaison[14].

Dans une approche distributionnelle, des indices des nuances sémantiques de mots sémantiquement proches peuvent aussi être repérés dans les contextes d'usage des mots en corpus. Ainsi, les graphes de proximité sémantique élaborés par Fabre *et al*. (1997) révèlent certaines facettes sémantiques des mots en fonction de leurs rapprochements avec d'autres mots du corpus ; facettes dont l'exploration demande un retour au corpus. Rossignol et Sébillot (2006) proposent, quant à eux, l'automatisation de la procédure de repérage de nuances sémantiques, sur la base de l'hypothèse que les contextes d'usage des mots sont porteurs d'informations concernant non seulement leurs points communs, mais aussi leurs différences sémantiques fines[15].

[13] Le modèle d'Edmonds et Hirst (*ibid*.) est en partie dépendant d'ontologies prédéfinies, tandis que les informations concernant les différences entre les quasi-synonymes sont modélisées à la main.
[14] Les auteurs proposent la comparaison séparée des **distinctions de dénotation**, des **distinctions expressives** et des **distinctions stylistiques**.
[15] Leur méthode vise plutôt à effectuer des rapprochements entre paires de mots distingués par une même nuance, sur la base de la notion d'analogie (Lepage, 2004).

2.2.3. La *différence* dans notre méthode

Au sein de notre travail, les EQVs clustérisés présentent des similarités aussi bien que des différences. La méthode d'estimation de la similarité permet d'avoir accès aux éléments qui rapprochent et qui distinguent les EQVs comparés, au niveau des langues source et cible, éléments représentés respectivement par leurs **contextes assimilateurs** et **dissimilateurs** (cf. 4.3.2.). Les similarités des EQVs au niveau de la LS permettent leur clustering, tandis que leurs différences au niveau de la LC conditionnent la sélection parmi les EQVs clustérisés pour la traduction.

Les EQVs clustérisés sont considérés comme substituables (à un degré plus ou moins élevé) en tant que traductions d'un sens du mot source, mais ce sont les différences au niveau de leur combinatoire au sein de la LC qui définissent leur **substituabilité effective** en contexte. La sélection lexicale est donc effectuée parmi les EQVs trouvés au sein du cluster choisi par le module de désambiguïsation comme représentant le sens d'une nouvelle instance du mot ambigu.

3. Similarité sémantique en contexte : la question de substituabilité

3.1. Substituabilité et similarité sémantique : quelle relation ?

Lorsque la similarité sémantique des mots est envisagée en contexte, une nouvelle notion entre en jeu : celle de **substituabilité** (ou de **permutabilité**). La substituabilité des mots est proportionnelle à leur similarité sémantique et peut être considérée comme un corollaire de cette relation. Elle est souvent regardée comme la possibilité d'utiliser alternativement des mots similaires dans le même contexte (de les permuter), sans changement du sens de l'énoncé. Elle peut pour autant constituer également un critère pour le repérage d'une relation entre deux mots ; leur similarité peut être considérée d'autant plus forte que la possibilité de leur substitution est grande.

L'utilisation de critères distributionnels pour l'estimation de la similarité des mots se retrouve déjà dans la discussion classique sur la synonymie. Là, ces critères sont utilisés de manière rigide, conformément au principe de substitution suivant : si un mot est substitué dans une phrase à un autre mot sans changer la **valeur de vérité** de la phrase, les deux mots ont la même **dénotation**[16]. Dans ce cas, les mots sont donc considérés comme des synonymes. Néanmoins, l'observation de situations et de contextes où la substitution d'expressions considérées comme référentiellement synonymes (ayant la même dénotation) pouvait provoquer le changement des valeurs de vérité de l'énoncé[17] a provoqué l'abandon de ce principe trop rigide de la logique extensionnelle. L'identité extensionnelle a cessé ainsi d'être considérée comme une condition suffisante garantissant la substituabilité, ce qui a justifié le recours aux logiques intensionnelles et à une sémantique des **mondes possibles**.

A part le ré-examen de la question de substituabilité *salva veritate*, la non-substituabilité d'expressions référentiellement synonymes dans certains contextes a constitué également l'argument classique pour rejeter l'idée même de synonymie (Fuchs, 1994 : 77) – dans le sens de synonymie absolue. Déjà Goodman (1952), souligne qu'il n'existe pas deux prédicats (même synonymes) qui soient remplaçables dans toute phrase sans provoquer un changement de valeur de vérité ; l'impossibilité de substitution dans un contexte de prédicats considérés comme ayant le même sens démontre immédiatement leur différence sémantique. La constatation que deux termes ont le même sens n'indique, pour Goodman, que leur degré de similarité sémantique est suffisant pour les buts d'un discours donné. Il serait donc préférable d'estimer des degrés ou des types de synonymie et de considérer ce degré comme degré d'**intersubstituabilité**.

Dans des travaux plus récents, la similarité sémantique des mots est considérée comme proportionnelle à leur **similarité distributionnelle**. La

[16] Le principe « *Eadem sunt qui substitui possunt salva veritate* » constitue un des principes classiques de la logique extensionnelle et est attribué à Leibniz.

[17] Un exemple classique est celui donné par Frege (1971 : 108), qui concerne les propositions : « l'étoile du matin est un corpus illuminé par le soleil » et « l'étoile du soir est un corps illuminé par le soleil ». Le remplacement d'un terme de la première proposition par un terme ayant la même dénotation mais un sens différent dans la deuxième (« étoile du matin » - « étoile du soir »), n'a aucune influence au niveau de la dénotation. Cependant, la pensée (qui constitue le contenu de la proposition) subit une modification : si quelqu'un ignorait que l'étoile du soir est l'étoile du matin, il pourrait tenir l'une de ces pensées pour vraie et l'autre pour fausse.

substituabilité concernant l'utilisation alternative des mots dans le même contexte, leur similarité distributionnelle peut être considérée dans des termes de substituabilité (Church *et al.*, 1994 ; Weeds, 2003). L'idée de substituabilité lexicale souligne ainsi la relation entre similarités sémantique et distributionnelle. Cette relation est centrale dans l'hypothèse distributionnelle de la similarité sémantique (Charles et Miller, 1989), où l'estimation de la parenté des mots est basée sur des critères distributionnels. Si la similarité sémantique est conçue de cette manière, les mots comparés peuvent être considérés comme substituables au sein des contextes qui définissent leur similarité. Le degré de leur substituabilité serait dépendant de leur similarité distributionnelle, de la même manière que le degré de leur similarité sémantique ; dans le cas de synonymes parfaits, la discrimination de leurs contextes serait impossible.

Cependant, ce principe de substituabilité (ou de permutabilité) des mots inclus dans une classe sémantique, comme celles construites par les méthodes distributionnelles, ne s'applique par toujours. Il se peut, en effet, que les mots d'une classe ne présentent pas systématiquement des comportements similaires. Pour répondre à la question de substituabilité, il est important de considérer ce qui constitue la spécificité des sens des mots rassemblés dans une classe. Car c'est la connaissance de ces nuances qui permet de définir plus précisément les **conditions** dans lesquelles une permutation est sémantiquement pertinente (Rossignol et Sébillot, 2006).

Pour Weeds (2003), la similarité distributionnelle paraît constituer un besoin plus « faible » que la similarité sémantique ; deux phrases peuvent être plausibles suite à une substitution sans avoir le même sens, mais elles ne peuvent pas avoir le même sens sans être toutes les deux plausibles. Autrement dit, la similarité distributionnelle n'implique pas toujours la similarité sémantique[18], tandis que la similarité sémantique implique plus souvent la similarité distributionnelle ; cette constatation démontre l'**asymétrie** de la relation entre similarité sémantique et distributionnelle.

Polguère (2003 : 122-123) utilise le terme **synonymes approximatifs** pour décrire des synonymes qui se distinguent aux niveaux du sens et de la

[18] Il est important de souligner que la similarité sémantique est considérée ici dans le sens de synonymie, parce que la similarité distributionnelle peut mette en évidence d'autres types de relations entre les mots.

combinatoire, ce qui empêche leur substitution dans tous les contextes. L'intersubstituabilité des mots n'est donc pas seulement une question de similarité sémantique, mais peut être également influencée par des facteurs caractérisant le texte où la substitution est supposée se produire ; ces facteurs concernent tant la cohésion et la cohérence du texte, que les conditions extralinguistiques dans lesquelles il est produit.

Plus précisément, Polguère (2003 :122-123) considère comme synonymes deux lexies L_1 et L_2 si leur remplacement dans une phrase fournit une nouvelle phrase à peu près équivalente sémantiquement, c'est-à-dire une **paraphrase approximative**. Même si les synonymes ne sont pas mutuellement substituables dans tous les contextes, il suffit d'être en mesure de trouver des contextes où la substitution paraphrastique est possible pour que le lien de synonymie soit établi. Ce besoin de considération d'un contexte large pour l'estimation de la substituabilité est déjà souligné par Quine (1953 : 56-57), d'après qui, la substitution de synonymes doit assurer la transformation des énoncés, dans lesquels elle se produit, en des énoncés synonymes d'une certaine manière (sans considération de préservation de leur valeur de vérité). Malgré la **circularité** de l'observation – les mots sont synonymes si leur substitution laisse leurs contextes synonymes – elle rend évident le besoin d'une notion de synonymie pour de longs segments de discours[19].

L'estimation de la pertinence d'une substitution n'implique pas seulement leur sémantique, mais la considération d'un ensemble d'autres facteurs. L'importance d'une telle estimation est grande, étant donné qu'elle pourrait constituer un critère pour l'estimation de la similarité lexicale (Miller et Charles, 1991 ; Polguère, 2003[20]) et qu'elle pourrait également empêcher la substitution inappropriée des mots dans les textes.

[19] Quine réduit la synonymie à l'aspect cognitif, puisque du point de vue de « la qualité poétique », la synonymie n'existe pas – les mots se distinguant, au moins, par « leur valeur poétique différente » (Quine, 1953 : 57).

[20] Polguère (2003 :122-123) considère comme synonymes deux lexies L_1 et L_2 si leur remplacement dans une phrase fournit une nouvelle phrase à peu près équivalente sémantiquement, c'est-à-dire une paraphrase approximative. Même si les synonymes ne sont pas mutuellement substituables dans tous les contextes, il suffit d'être en mesure de trouver des contextes où la substitution paraphrastique est possible pour que le lien de synonymie soit établi.

3.2. Asymétrie de la substituabilité

Une autre question importante autour de la notion de substituabilité concerne son caractère symétrique : est-ce qu'il est toujours possible de substituer un élément B à un élément A s'il a été démontré que A est substituable à B ?

En effet, Weeds (2003 : 17) souligne que le concept de la substituabilité présente un caractère hautement **asymétrique**. Il est possible d'estimer l'adéquation de substitution du mot B au mot A, indépendamment de l'estimation de l'adéquation de substitution du mot A au mot B[21]. Cette non intersubstituabilité peut être observée, par exemple, dans le cas d'une relation d'hypéronymie-hyponymie : un hypéronyme (par ex. *animal*) peut être plus facilement utilisé à la place d'un de ses hyponymes (par ex. *dog*), qu'un hyponyme à la place de son hyperonyme. Weeds explore la possibilité d'exploitation de l'asymétrie inhérente à la substitution pour proposer une mesure asymétrique de la similarité distributionnelle, étant donné que cette similarité est souvent définie en des termes de substituabilité.

La question de substituabilité touche aussi la problématique autour de la paraphrase (Fuchs, 1994 : 77). Là aussi, si une relation de paraphrase est symétrique, les éléments liés doivent pouvoir commuter entre eux dans n'importe quel type de contexte. Les paraphrases ne sont pourtant mutuellement substituables qu'en faisant abstraction des différences de sens, en ne considérant que l'invariant. La substitution peut donc devenir difficile ou même impossible dans des contextes où les différences sémantiques se trouvent mises en avant.

3.3. La substituabilité dans un cadre de traduction

Des mots qui semblent substituables sur la base de certains critères formels (comme la similarité distributionnelle) peuvent ne pas l'être vraiment dans un texte réel à cause d'autres paramètres qui empêchent leur substitution. Le nombre de ces paramètres augmente dans un cadre de traduction, où la

[21] Ceci explique, d'après Weeds (*ibid.*), pourquoi l'hypéronyme d'un mot nous vient facilement à l'esprit lorsque nous cherchons un mot lui étant sémantiquement lié (par ex. *animal* pour *dog*), tandis que l'inverse (la réponse de *dog* comme mot similaire à *animal*) est plus difficile.

substituabilité peut concerner les EQVs des mots source dans la LC. Là des besoins supplémentaires apparaissent, comme la préservation du sens du mot source dans la traduction et la pertinence du texte créé dans la LC. Il se peut donc que deux EQVs sémantiquement apparentés soient considérés comme substituables en tant que traductions d'un mot source en contexte, s'ils peuvent tous les deux exprimer son sens dans la LC, mais qu'ils ne soient pas pour autant tous les deux adéquats au sein du texte de la LC.

Dans le cadre de notre travail, la question de substituabilité concerne les EQVs clustérisés des mots ambigus, qui sont sémantiquement liés. La notion de substituabilité peut être conçue de deux manières :

(1) Premièrement, il serait possible de parler de substituabilité de deux EQVs en tant que traductions d'une nouvelle instance du mot source en contexte. La similarité des EQVs étant induite par la similarité distributionnelle de leurs contextes source, leur substituabilité concerne la possibilité de leur utilisation alternative en tant que traductions de l'instance du mot source, lorsque son contexte est proche de celui ayant induit leur similarité.

(2) Deuxièmement, la similarité des EQVs pouvant également être estimée par référence à leur similarité distributionnelle dans la LC, leur intersubstituabilité pourrait être considérée par rapport à ce type de contexte. Dans un cadre de TA, les contextes cible pourraient être composés des traductions des autres mots de la phrase d'entrée.

Ces deux aspects de la substituabilité seront développés davantage dans le paragraphe suivant.

4. Similarité sémantique et sélection lexicale

4.1. Sélection entre mots similaires dans un cadre monolingue

Les conditions d'utilisation de mots qui présentent des différences sémantiques frappantes sont assez claires pour être modélisées et efficacement exploitées pour la sélection lexicale dans les applications du TAL. Ces différences sont souvent reflétées dans la distribution des mots, ce qui permet le repérage facile des indices contextuels permettant leur identification. Ces informations peuvent également être aisément décrites à l'aide d'ontologies ou de réseaux sémantiques.

En revanche, les critères de sélection dans le cas de mots présentant des différences sémantiques fines ne sont pas si évidents ni aussi facilement repérables, et ils se prêtent moins aisément à la modélisation. Ces critères peuvent être liés à la sémantique des mots en question, c'est-à-dire aux éléments qui les différencient comme les **idées accessoires** ou les **nuances de sens** (Edmonds, 1997, 1999 ; Edmonds et Hirst, 2002), mais cela n'est pas toujours le cas.

Reiter et Sripada (2004), qui travaillent dans le domaine de la Génération Automatique de Textes (GAT), défendent l'idée que la sélection parmi des quasi-synonymes est plutôt déterminée par des facteurs non sémantiques, comme les préférences, le style d'écriture et l'idiolecte des auteurs (dans le sens des différences individuelles dans l'usage de la langue) ; elle est également dépendante du facteur de « sécurité », qui concerne la sélection de mots « neutres » (ayant moins de connotations) pour éviter des implications indésirables. Ces facteurs ont une portée sur les collocations, la répétition des mots dans le texte et leur position.

Pour Stede (1999 : 19-20), les facteurs extra-linguistiques qui influencent la sélection lexicale concernent l'utilisation de mots différents par les locuteurs pour dire approximativement la même chose dans des situations différentes. Les critères de sélection peuvent être liés, dans ce cas, aux genres différents, au style (formel ou familier) ou au besoin d'utiliser des mots marqués, indiquant une attitude vers une situation. Ces critères peuvent être regroupés dans trois

catégories : les **contraintes collocationnelles**, où le choix d'un mot influence celui d'un autre ; les **contraintes de focalisation**, qui confèrent une saillance relative à certaines parties d'un texte en influençant ainsi les choix lexicaux, et les **contraintes pragmatiques**. Ces dernières modélisent la relation entre l'intention du locuteur de poursuivre certains buts communicatifs et les moyens linguistiques qu'il utilise, dont la sélection lexicale. Cependant, à l'exception des collocations, qui peuvent être modélisées en tant que contraintes, les autres facteurs de sélection lexicale seraient mieux considérés comme des **préférences**. Selon Stede, la représentation des connotations au niveau du lexique dans un système de GAT et l'inclusion d'informations pragmatiques dans l'entrée du générateur provoqueraient des conflits. Les buts stylistiques particulaires peuvent être ou ne pas être réalisés et les choix lexicaux peuvent être faits en adoptant des perspectives différentes.

4.2. Sélection entre mots similaires dans un cadre de Traduction

Dans le domaine de la Traduction, derrière la problématique de la sélection lexicale se pose la question de la lexicalisation des sens dans des langues différentes (cf. §2.3, chapitre 1). Dans ce cadre, la sélection lexicale consiste à choisir, parmi les candidats de traduction d'un mot source, l'équivalent le plus adéquat pour une nouvelle instance de ce mot. La difficulté de cette tâche dépend beaucoup de la sémantique du mot source, mais aussi des relations sémantiques entre ses EQVs de traduction[22]. La complexité de la sélection lexicale augmente en allant des mots présentant de l'ambiguïté contrastive, où des distinctions de sens claires sont impliquées, aux mots polysémiques ou à sens vague, où les distinctions sémantiques ne sont pas toujours évidentes.

Dans le cas d'**ambiguïté contrastive**, le choix entre les candidats de traduction peut se faire facilement à l'aide du contexte de la nouvelle instance du mot source, ou même à l'aide de connaissances du domaine. Dans le cadre de la traduction manuelle, nous pourrions même dire que dans certains cas un des

[22] Nous faisons ici abstraction des facteurs extra-linguistiques qui influencent la sélection et nous nous focalisons à ces aspects qui ont une portée sur la sélection lexicale, qui se manifestent au sein des textes et qui peuvent être identifiés à partir de corpus.

candidats s'impose de telle manière au traducteur, que l'autre n'arrive même pas à traverser son esprit. Ce qui facilite la sélection est que les équivalents sont antagonistes (comme d'ailleurs les sens du mot source), l'utilisation de l'un excluant l'utilisation de l'autre. La sélection parmi les candidats de traduction est beaucoup plus difficile dans le cas où ils ont une **relation sémantique proche**. Dans les cas des (quasi-) synonymes, les critères de sélection ne sont pas toujours clairs et l'utilisation d'équivalents différents peut paraître possible dans un contexte donné.

La sélection entre les candidats de traduction peut être effectuée en prenant, d'abord, en considération l'usage du mot source en contexte et ses nuances sémantiques et, dans un deuxième temps, les différences de sens entre les équivalents dans la LC. Dans un système de TA, le repérage des nuances du mot source dans un contexte donné peut être fait pendant la phase d'analyse, tandis que la sélection de l'équivalent de traduction exprimant les mêmes nuances dans la LC peut être effectuée pendant la phase de sélection lexicale. Ainsi, les possibilités de ce qui est exprimé dans le texte de la LS sont déterminées pendant l'analyse ; elles constituent des **préférences** pour la sélection et conditionnent ensuite la sélection des mots adéquats dans la LC. La sélection lexicale consiste donc à sélectionner les mots permettant de satisfaire un ensemble de **préférences** (probablement conflictuelles), afin d'exprimer certaines nuances, établir le style souhaité et respecter les contraintes collocationnelles et syntaxiques.

Le processus de sélection lexicale est divisé en deux phases de ce type dans le modèle proposé par Edmonds et Hirst (2002 ; Edmonds, 1997, 1999) :

a. la sélection entre des options illustrant des différences sémantiques de granularité grossière. Ces options sont représentées par des clusters de mots similaires (quasi-synonymes), qui se trouvent en conflit.

b. la sélection d'un quasi-synonyme du cluster choisi, voire la sélection entre des options présentant des différences sémantiques et stylistiques fines.

Edmonds et Hirst considèrent la quasi-synonymie comme la norme dans le transfert lexical en traduction. La traduction exacte est pour eux probablement

impossible, étant donné que chaque possibilité de traduction peut omettre quelque nuance ou exprimer une nuance indésirable.

Le problème de la sélection lexicale se pose également dans le cadre de l'aide à la rédaction dans une deuxième langue (Tiedemann, 2001), où il concerne la prédiction du mot le plus correct en contexte. Dans le modèle de Tiedemann, les (pre- et post-) contextes des deux langues sont exploités. Les candidats de traduction sont fournis dans des bases de données terminologiques bilingues, automatiquement extraites à partir de corpus parallèles alignés, et la prédiction du mot le plus adéquat en contexte est effectuée par une analyse du contexte présent dans les phrases alignées du corpus. Les prédictions du système dépendent à chaque moment du contexte disponible et l'acceptation, ou le rejet, des prédictions en question se fait par l'utilisateur de manière interactive.

En effet, le problème de la sélection lexicale dans un cadre de traduction pourrait être illustré à l'aide d'un **vide** dans la traduction, qui devrait être rempli par un des candidats de traduction du mot source. Cette approche est adoptée par Vickrey *et al*. (2005), qui considèrent le **remplissage de blancs** (*blank filling*) comme un **problème simplifié de traduction**. Notre méthode de sélection lexicale adopte une approche similaire à celle de Vickrey *et al*.

4.3. Prise en compte de la substituabilité par notre méthode de sélection lexicale

4.3.1. Filtrage des résultats de la désambiguïsation

Notre processus de sélection lexicale se base initialement sur les résultats de la méthode de désambiguïsation ; cette méthode exploite le contexte source des nouvelles instances des mots ambigus et arrive à restreindre le nombre de leurs candidats de traduction possibles. Dans certains cas, la sortie de la désambiguïsation concerne un seul EQV du mot source, et le processus de désambiguïsation coïncide ainsi avec la sélection lexicale. Cependant, dans le cas où la sortie de la désambiguïsation concerne des possibilités de traduction multiples, l'ensemble d'EQVs proposé doit être filtré.

Dans un cadre semi-automatique, ce filtrage pourrait être effectué par l'utilisateur. Dans un cadre automatique, où l'intervention humaine n'est pas possible, un processus de sélection lexicale permettant de choisir le plus adéquat des équivalents proposés doit être mis en place. La méthode de sélection lexicale que nous proposons prend en compte des informations venant des contextes lexicaux de la LC.

Le fonctionnement de cette méthode est basé sur l'idée que les EQVs sémantiquement similaires proposés par le module de désambiguïsation peuvent ne pas être tous adéquats et librement substituables au sein de la traduction. En effet, les relations de similarité des EQVs exploitées pour la désambiguïsation sont établies par référence aux contextes source des mots ambigus. Ainsi, même si les EQVs paraissent adéquats pour traduire le mot source dans le nouveau contexte, il se peut que leur utilisation ne soit pas toujours « naturelle » dans la LC. Même si les mots ont un noyau sémantique commun, il est possible que leur utilisation dans le texte cible ne soit pas adéquate, soit à cause de leurs idées accessoires et de leurs connotations, soit en raison de la structure du texte de la LC.

Le contexte exploité par notre méthode de sélection de traduction est le contexte lexical proche (au sein de la même unité de traduction). Des facteurs extralinguistiques pouvant avoir une portée sur ce processus ou des paramètres repérables au niveau du document ne seront pas considérés.

Les méthodes de désambiguïsation et de sélection lexicale proposées seront présentées en détail dans le chapitre suivant. Par la suite, nous présenterons les informations qualitatives concernant les EQVs clustérisés qui sont fournies par le calcul de similarité et qui peuvent servir à décrire leur possibilité de substitution.

4.3.2. Contextes assimilateurs et dissimilateurs

4.3.2.1. Contextes assimilateurs

Le calcul de similarité sémantique effectué (cf. §2.3, chapitre 6), met en évidence les traits contextuels qui rapprochent et qui distinguent les EQVs comparés. Les contextes lexicaux qui rapprochent deux EQVs et révèlent leur

similarité sont constitués des traits qui leur sont communs. L'ensemble de ces traits peut être caractérisé comme les **contextes assimilateurs** des EQVs et peuvent servir à décrire leur relation de similarité. En effet, l'observation de ces traits illustre les aspects de la sémantique des EQVs par rapport auxquels ils sont proches.

Le terme « co-textes assimilateurs »[23] est utilisé par Fuchs (1994 : 134-141) pour décrire les contextes linguistiques qui permettent l'établissement d'une relation de parenté dans le cas de **paraphrases**. Selon la conception dynamique de la signification, les paraphrases ne partagent qu'un **air de famille**, ce qui signifie qu'elles sont reliées par des relations sémantiques locales, de type associatif, construites par le jeu de l'interprétation.

> « *S'agissant de la parenté sémantique susceptible de fonder une relation de* ***paraphrase*** *entre énoncés, la situation semble bien être celle que décrit Wittgenstein, à savoir des* ***similitudes*** *plus ou moins locales construites résultativement à l'issue de l'interprétation des énoncés, et non pas un noyau commun partagé par tous les énoncés interparaphrastiques (…) La signification globale d'un énoncé résulte de l'interaction des marqueurs qui le composent ; or ces marqueurs correspondent à des* ***opérateurs différents*** *à partir desquels, par des* ***cheminements différents****, peuvent être construites résultativement certaines valeurs : parler de parenté, c'est être en mesure de considérer que les valeurs ainsi construites en co-texte se recoupent, se chevauchent*. » (*ibid*. : 133-134).

Un co-texte **assimilateur** est donc, dans le cas de la paraphrase, un co-texte qui permet « *de construire, à partir de deux marqueurs (opérateurs) différents, des valeurs qui se recoupent plus ou moins fortement* ». (*ibid*. : 134).

Dans notre travail, deux EQVs similaires sont considérés comme potentiellement intersubstituables au sein de leurs contextes assimilateurs. L'intersubstituabilité d'une paire d'EQVs est dépendante du degré de leur similarité sémantique, degré quantifiable à l'aide du score de similarité qui lui est attribué. Même dans le cas d'une similarité forte, deux EQVs sont rarement des synonymes absolus substituables dans tous les contextes. Il existe presque toujours des différences, pouvant même empêcher cette substitution.

[23] Le terme « co-texte » est utilisé par Fuchs pour décrire le contexte linguistique, qui doit être différencié du « contexte » qui englobe ce qui est extérieur du langage tout en faisant partie de la situation d'énonciation. N'exploitant pas le contexte situationnel, dans ce travail nous utilisons toujours le terme contexte pour décrire le contexte linguistique.

4.3.2.2. Contextes dissimilateurs

Certaines des différences existantes entre mots sémantiquement similaires peuvent être représentées à l'aide des traits contextuels qui les distinguent au niveau de la distribution. De tels traits existent toujours et se mettent en contraste avec les traits contextuels assimilateurs des mots ; plus précisément, il s'agit des traits qui sont spécifiques aux mots comparés, dont le nombre peut être plus ou moins élevé et la pertinence variable. Ces traits, qui empêchent le rapprochement des mots comparés et qui montrent leur différence, peuvent servir également à la description de leur relation de similarité. Nous appelons (à la suite de Fuchs, 1994 : 138) ces contextes des **contextes dissimilateurs**.

Les traits de ces contextes décrivent la manière dont les mots se différencient du point de vue de leur distribution lexicale ; ils peuvent ainsi servir à l'identification des contextes où leur permutation n'est pas possible. L'observation de ces traits nous permet de comprendre sous quels aspects de leur sémantique les EQVs sont éloignés. Sur la base de l'hypothèse contextuelle de la similarité sémantique, les ensembles de traits distinctifs peuvent être considérés comme démontrant la dissimilarité sémantique des mots, ou même illustrant des sens ou des sous-sens exprimés exclusivement par certains des mots comparés. La dissimilarité de deux mots pourrait donc être considérée comme proportionnelle au nombre de ces traits. En outre, certains traits distinctifs seraient plus pertinents que d'autres lors de l'estimation de la dissimilarité des mots, comme dans le cas du calcul de similarité sémantique où certains traits sont considérés comme plus pertinents pour l'estimation de leur similarité. L'estimation de l'importance des traits pour les jugements de dissimilarité peut se faire à l'aide des poids qui leur sont attribués lors du calcul du coefficient de Jaccard.

Les contextes assimilateurs et dissimilateurs d'une paire d'EQVs clustérisés peuvent être fournis optionnellement à la sortie du processus de clustering. Ces traits peuvent servir à la description de la relation de similarité (ou de dissimilarité) entre les EQVs en question, si un tel besoin de description des sens apparaît. En outre, ces ensembles de traits peuvent être exploitées lors des étapes

de désambiguïsation et de prédiction de traduction utilisant les résultats de l'acquisition de sens (cf. chapitre 8).

4.3.3. Substituabilité des équivalents clustérisés

4.3.3.1. Substituabilité des équivalents par rapport aux contextes source

La sélection d'un parmi les EQVs possibles d'un mot source n'est pas toujours imposée par des contraintes strictes. La rigidité de ces contraintes et leur nature même, dépendent des relations de similarité existant entre les EQVs et des distinctions sémantiques induites par les résultats du clustering sur le mot source.

Dans le cas de sens antagonistes, par exemple, proposés par des clusters distincts, les contraintes sémantiques qui s'imposent concernant la sélection de traduction sont fortes. En revanche, dans le cas de distinctions sémantiques moins frappantes, comme c'est le cas de clusters chevauchants, la sélection entre les EQVs peut être conçue comme plus « flexible ». Dans ce cas, il serait possible de parler de **préférences** plutôt que de **contraintes**. Cette manière d'envisager la sélection entre EQVs s'applique encore plus dans le cas d'EQVs clustérisés exprimant le même sens du mot source et présentant des relations sémantiques proches, relations qui permettent une plus grande flexibilité au niveau de la sélection.

Les contextes assimilateurs et dissimilateurs des EQVs mis en évidence par le calcul de similarité au niveau de la LS peuvent servir à décrire leur substituabilité par rapport aux contextes de la LS. En effet, nous pouvons dire qu'au niveau de la LS, c'est plutôt les contextes assimilateurs qui nous intéressent. Il s'agit des contextes qui révèlent la similarité entre les EQVs et qui permettent le repérage des sens du mot source.

Les éléments d'un cluster donné pourraient être considérés (plus ou moins) intersubstituables comme traductions du sens décrit par le cluster. Leur intersubstituabilité est en effet proportionnelle à leur similarité. La similarité entre deux EQVs peut être décrite à l'aide des informations contextuelles qui la

révèlent, c'est-à-dire par leurs contextes assimilateurs[24]. Par exemple, dans le cas des EQVs *μονάδα* (monada) et *εγκατάσταση* (egkatastasi), qui ont la relation de similarité la plus forte parmi les EQVs de *plant* (0,209, 0,177)[25], quelques éléments des contextes anglais qui leur sont communs sont les suivants : {*product, air, exceed, sulphur, limit, combustion, emission, industrial, dioxide, follow, apply, way*, …}. Cet ensemble d'éléments nous donne des informations sur la sémantique des deux EQVs : ils traduisent *plant* quand il apparaît dans des contextes où il est question de l'impact négatif des émissions d'usines de combustion sur l'environnement. Les deux EQVs pourraient donc être considérés comme éventuellement interchangeables au sein de tels contextes[26].

Les contextes dissimilateurs des EQVs, qui sont mis en évidence aussi par le calcul de similarité, décrivent la manière dont les EQVs se différencient du point de vue de leur distribution lexicale et peuvent servir à l'identification des contextes qui empêchent leur permutation. Certains des traits des contextes assimilateurs et dissimilateurs de chaque paire d'EQVs clustérisés de *plant* sont décrits dans la figure 1. Les ensembles précédés par le symbole '=' contiennent les traits communs aux EQVs de la paire, tandis que les ensembles précédés par '≠ *EQV* ' décrivent les traits qui caractérisent uniquement l'EQV en question.

[24] Cf. §4.3.2, chapitre 7.

[25] Le premier score donné entre parenthèses pour une paire d'équivalents correspond au score selon les contextes anglais et le deuxième au score selon les contextes grecs, les deux dans l'ensemble du corpus.

[26] Si un besoin de définition des sens décrits par les clusters apparaît (par exemple, dans un cadre de création de dictionnaires destinés à des humains), les contextes assimilateurs des EQVs clustérisés pourraient être exploités dans ce but. Cependant, la description des sens véhiculés par un mot n'est pas toujours nécessaire dans un cadre de TAL. Selon Ide *et al*. (2002), la plupart des applications ne nécessitent pas de connaissances sur la nature des sens, mais seulement des informations relatives à la distinction des instances d'un mot ambigu en fonction du sens dans lequel elles sont utilisées.

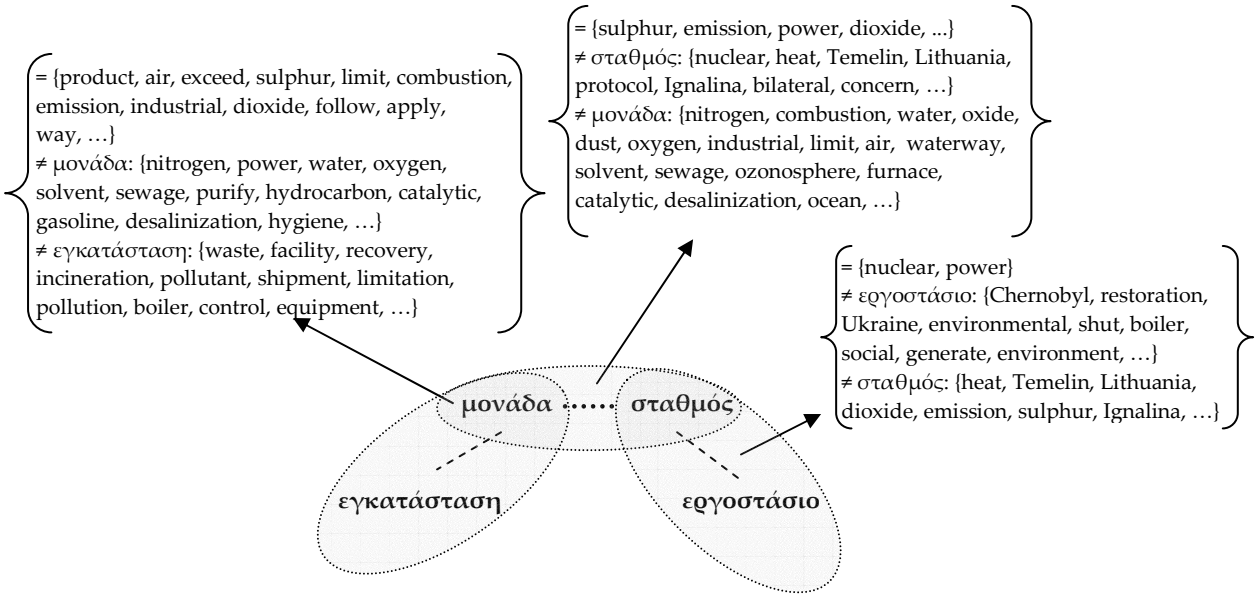


Figure 1. Contextes assimilateurs et dissimilateurs source des EQVs de *plant*

Bien évidemment, il se peut que les traits qui distinguent deux EQVs, les rapprochent d'autres. Par exemple, les traits dissimilateurs de *σταθμός* par rapport à *εργοστάσιο* : {*sulphur*, *dioxide*, *emission*}, font partie de ses traits assimilateurs par rapport à *μονάδα*.

4.3.3.2. Substituabilité des équivalents par rapport aux contextes cible

Des contextes assimilateurs et dissimilateurs sont aussi révélés au niveau de la LC, par le calcul de similarité effectué sur les contextes de ce type. Des traits de ces contextes pour les EQVs clustérisés de *plant* sont donnés dans la figure 2. Au niveau de la LC, nous sommes surtout intéressés par les contextes dissimilateurs des EQVs. Ces contextes, qui décrivent les différences au niveau de l'usage dans la LC entre EQVs similaires, peuvent constituer une sorte de « filtre » lors de la sélection lexicale.

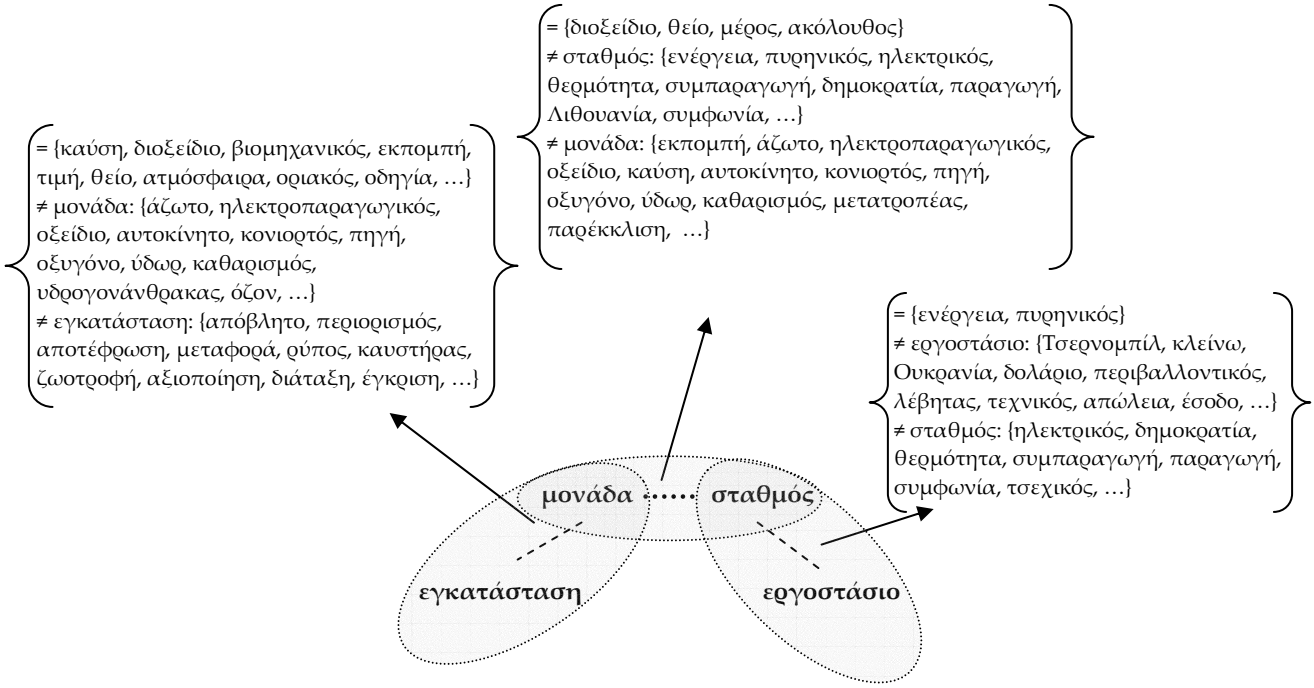


Figure 2. Contextes assimilateurs et dissimilateurs cible des EQVs de *plant*

Nous rappelons que la sortie du module de désambiguïsation peut être constituée par un ensemble d'EQVs sémantiquement similaires formant un cluster. Dans un cadre de Traduction Assistée par Ordinateur, ces EQVs peuvent constituer des propositions au niveau de mots, la sélection finale étant effectuée par l'utilisateur.

Dans un cadre automatique, cette sélection peut être faite par prise en compte du contexte de la LC. Nous avons déjà souligné que les informations les plus pertinentes pour la sélection entre mots présentant des distinctions sémantiques fines se situent au niveau de leurs différences. Par conséquent, la sélection de l'EQV le plus adéquat à la traduction, parmi les EQVs clustérisés, doit se faire par rapport aux différences des EQVs au niveau de la LC. Ces différences sont repérées au niveau de leur distribution et sont reflétées dans leurs ensembles de traits dissimilateurs dans la LC.

La possibilité d'avoir accès à des informations permettant d'affiner les relations sémantiques entre les EQVs, en prenant en compte leurs différences, permet de caractériser notre approche d'approche à **profondeur variable** (Kayser et Coulon, 1981). Un traitement à profondeur variable utilise une description progressive des données, exploite des stratégies différentes en fonction de la

qualité des résultats souhaités et contrôle cette qualité continuellement [27]. Ce type de traitement s'appuie sur l'hypothèse que l'exactitude de l'analyse computationnelle doit pouvoir varier et même être approfondie, si un tel besoin se présente. Chaque mot doit donc être lié à une quantité de connaissances ordonnée, disponible de manière progressive lorsque la profondeur de l'analyse augmente.

Cette question de la profondeur touche ici à celles de la quantité et de la nature des informations prises en compte pour les différentes tâches. Selon la tâche envisagée, la quantité et la nature des informations nécessaires divergent. Ainsi, les méthodes de WSD et de sélection lexicale, ayant des besoins divergents quant aux informations exploitées pour accomplir les tâches en question, les informations utiles seront donc différentes. Si les traits assimilateurs de la LS, qui permettent de capter le noyau sémantique commun des EQVs, sont suffisants pour l'étape de WSD, la sélection lexicale requiert, elle, la prise en compte des différences entre les EQVs, reflétées dans leurs traits dissimilateurs de la LC. Ce sont donc ces informations qui seront utilisées lors de l'étape de sélection lexicale.

4.3.3.3. Contraintes vs. préférences de sélection

Les informations qui servent au clustering et à la modélisation des correspondances entre sens et EQVs, et qui sont exploitées par la suite pour la désambiguïsation et la sélection lexicale, sont extraites à partir de notre corpus d'apprentissage. La méthode de sélection lexicale parvient à choisir parmi un ensemble d'EQVs illustrant le sens de la nouvelle instance du mot source – qui constitue la sortie du module de désambiguïsation – le plus approprié pour traduire l'instance en question. Nous avons voulu vérifier si l'élimination des autres EQVs de cet ensemble par le module de sélection lexicale est toujours sémantiquement motivée et quel est l'impact des limitations inhérentes au travail sur corpus sur la sélection effectuée.

[27] Les représentations de connaissances standard de l'Intelligence Artificielle se situent pourtant à un niveau de profondeur fixe, ce qui signifie que les objets manipulés sont décrits par une quantité d'informations qui reste constante pour chaque tâche.

Nous avons donc examiné si la proposition d'un EQV est toujours sémantiquement motivée ou si elle est due aux informations retenues pendant l'apprentissage. Pour faire cela, nous avons consulté une autre ressource, le corpus EUROPARL[28], que nous avons fouillé pour repérer des occurrences des EQVs au sein des contextes similaires à ceux dont ils sont exclus. La consultation de cette ressource a effectivement démontré la possibilité d'utilisation des EQVs dans des contextes où ils n'apparaissent pas dans notre corpus d'apprentissage.

Nous donnons un exemple : dans le cas du mot *plant*, une relation sémantique forte est repérée entre les EQVs *μονάδα* (monada) et *εγκατάσταση* (egkatastasi). En observant leurs contextes dissimilateurs, nous constatons que *μονάδα* traduit des usages de *plant* qui font référence à des *unités* de dessalement de l'eau et à des *stations* d'épuration des eaux usées, usages pour lesquels *εγκατάσταση* n'est pas utilisé dans le corpus d'apprentissage. Cependant, dans le corpus EUROPARL nous trouvons des instances des termes complexes *sewage treatment plant* et *desalinisation plant* où l'EQV *εγκατάσταση* est utilisé : ***εγκατάσταση*** *καθαρισμού υδάτων* / ***εγκατάσταση*** *βιολογικού καθαρισμού (των λυμάτων)* et ***εγκατάσταση*** *αφαλάτωσης*. L'utilisation de cet EQV ne devrait donc pas être complètement exclue dans un tel contexte.

La découverte de ces contre-exemples a confirmé notre intuition selon laquelle dans les cas de distinctions sémantiques fines, l'adoption de contraintes strictes concernant la substitution des EQVs ne serait pas justifiée. Dans le cas de telles distinctions, il serait mieux de déterminer des **préférences** concernant l'utilisation des EQVs. Ces préférences permettraient la proposition d'EQVs plus ou moins substituables au sein du nouveau contexte, accompagnés d'un taux de **confiance** ; cette confiance pourrait correspondre à la somme des poids des traits d'un EQV qui sont repérés dans le nouveau contexte. Ces préférences pourraient être perçues comme des **contraintes relatives**, se situant à mi-chemin entre le libre choix et les contraintes linguistiques (Fuchs, 1994 : 147)[29]. Dans le cas

[28] Ce corpus nous sert de corpus d'évaluation pour les méthodes de désambiguïsation et de sélection lexicale.

[29] La langue peut imposer ou privilégier une formulation donnée sans pourtant exclure les autres, ou laisser ouverte toute une gamme de formulations. Dans le cas de la paraphrase, si les formulations alternatives sont équi-possibles du point de vue de la langue, le choix final ne reflète que les préférences du sujet. Cependant, il se peut que les contraintes linguistiques imposent à l'individu une formulation donnée à l'exclusion de toute autre. Les contraintes relatives guident l'énonciateur préférentiellement vers le choix d'une formulation donnée et il peut s'y conformer ou

d'enrichissement du corpus avec de nouveaux textes, les informations concernant les contextes assimilateurs et dissimilateurs des EQVs qui permettent l'établissement de préférences, pourraient être mises à jour.

Il semble que cette manière de concevoir le libre choix et les contraintes imposées par la langue présente des avantages par rapport à une conception rigide, dans le cas de mots fortement liés. Une telle conception des notions de préférence et de choix est, d'après Stede (1999 : 33), tout à fait pertinente dans le cadre de la traduction et de la génération multilingue. Des paramètres contextuels en guise de régularités doivent être dégagés, privilégiant une solution par rapport à une autre ou permettant le libre choix, pour la modélisation des préférences et la sélection. Selon Stede, les critères utilisés peuvent être, syntaxiques – concernant les contextes qui acceptent ou refusent les mots – ou sémantiques – concernant leurs propriétés sémantiques. Contrairement à une situation de communication entre humains, où les sujets ont des connaissances concernant leurs marges de choix entre énoncés sémantiquement apparentés, une bonne description de ce type de connaissances est très importante dans un cadre de traitement automatique, pour modéliser les possibilités d'utilisation d'unités différentes selon les contextes.

CONCLUSION

Dans ce chapitre, nous avons analysé la notion de **similarité sémantique**, notion centrale dans notre méthode d'acquisition de sens, mais également exploitée par les méthodes de désambiguïsation et de sélection lexicale que nous proposerons. Nous avons étudié la manière dont cette notion est conçue dans un cadre cognitif, ainsi que les approches d'estimation de la similarité sémantique utilisées dans ce cadre. Nous avons ensuite regardé la manière dont cette notion est prise en considération dans un cadre de traitement automatique et avons

non ; cela peut conduire à des formulations stylistiquement « marquées ». Les préférences stylistiques de l'énonciateur se manifestent encore plus clairement lorsque le choix est laissé libre par une absence de contraintes linguistiques (Fuchs, 1994 : 158). Les énoncés produits sont donc le résultat d'une série de décisions électives, qui concernent un choix entre les formes disponibles, qui, le plus souvent, ne sont pas équivalentes à tous égards.

analysé des différences observées au niveau du traitement, en fonction de la force et de la nature des relations de similarité des mots.

En faisant le lien entre la similarité sémantique des mots et leur substituabilité, l'importance de la relation pour la sélection lexicale, dans un cadre monolingue et bilingue, a été mise en évidence. Dans le dernier paragraphe de ce chapitre, nous avons analysé la manière dont les relations de similarité entre les mots sont prises en compte par notre méthode de sélection lexicale, dans le but de choisir l'équivalent de traduction le plus adéquat des mots source en contexte. Le fonctionnement des méthodes de désambiguïsation et de sélection lexicale sera davantage analysé et illustré par des exemples issus de notre travail dans le chapitre 8.

CHAPITRE 8

DESAMBIGUÏSATION POUR LA SELECTION LEXICALE

INTRODUCTION

Pour Levý (1967), la traduction humaine est un processus de décision, qui peut être décrit dans les termes de la théorie du jeu : l'ensemble des solutions possibles constitue le **paradigme** et le **choix** entre elles n'est pas aléatoire (dans la mesure où les alternatives ne sont pas équivalentes), mais lié au contexte. La traduction étant en même temps interprétation et génération, les processus de décision opératoires dans ce cadre sont de deux types : dans un premier temps, choisir entre les éléments du paradigme sémantique du mot au sein du texte source et, dans un deuxième temps, choisir à partir du paradigme des mots de la LC, celui qui correspond plus ou moins au sens sélectionné lors de la première étape. Le choix d'une unité lexicale est gouverné par un système d'**instructions**

conscientes et **inconscientes**, **objectives** et **subjectives**[1] ; le symbole terminal utilisé dans la traduction peut donc être étudié par rapport au système d'instructions responsable de son apparition.[2]

Un corpus parallèle permet de reconstruire le système d'instructions objectives, c'est-à-dire dépendantes du matériel linguistique, qui sont responsables du choix d'une traduction d'une instance d'un mot source au sein d'un contexte précis. Ces instructions peuvent donc être exploitées, tout d'abord, pour choisir le sens de l'instance du mot source en contexte (à partir de son paradigme sémantique) et, ensuite, pour aboutir à une décision concernant sa traduction dans la LC (à partir du paradigme de ces traductions possibles).

Dans un cadre automatique, la première étape, celle de la désambiguïsation, est parfois esquivée car considérée comme non nécessaire à la sélection de la traduction correcte. Le débat autour de la nécessité de cette étape est assez vif. Dans les premiers paragraphes de ce chapitre, nous présenterons les principales positions sur le sujet, en référence aux systèmes de TA statistique (*Statistical Machine Translation* (SMT)). Ensuite nous proposerons une méthode de désambiguïsation lexicale (*Word Sense Disambiguation* (WSD)) qui exploite les résultats de notre méthode d'acquisition de sens, ainsi qu'une méthode de sélection lexicale, qui se fonde sur les résultats de la WSD afin de sélectionner la traduction la plus adéquate d'un mot source en contexte.

1. Besoin de désambiguïsation pour la Traduction Automatique ?

1.1. Distance inter-langue et ambiguïté parallèle

La réponse intuitive et « évidente » à la question du besoin de désambiguïsation dans la TA considère que la désambiguïsation constitue une étape importante du processus de traduction et que des modules appropriés devraient être intégrés dans les systèmes de TA. Cependant, le phénomène

[1] Dans le cadre de la traduction manuelle, les instructions subjectives sont dépendantes de la structure de la mémoire du traducteur, de ses standards esthétiques, etc.
[2] Pour Levý, il serait possible de construire un « modèle génératif » de la traduction, en employant les méthodes utilisées pour la définition des problèmes de décision.

d'**ambiguïtés parallèles** (Resnik et Yarowsky, 2000 ; Altenberg et Granger, 2002b) observé entre les mots de deux langues, suscite parfois des doutes quant à la nécessité d'une étape de désambiguïsation dans les applications de TAL relatives à la Traduction : lorsque l'équivalent de traduction présente la même ambiguïté que le mot source, la traduction de ce mot peut se faire directement et ne nécessite donc pas l'identification du sens précis véhiculé par le mot.

Le degré d'ambiguïté parallèle dépendant fortement de la **distance** entre les langues impliquées dans le processus de traduction[3], ce paramètre a un impact important sur le besoin de WSD. Resnik et Yarowsky (*ibid*.) ont mené une série d'expériences qui visent à évaluer cet impact. Les résultats obtenus démontrent que le besoin de WSD diminue entre langues relativement proches, par rapport à des langues plus distantes.

Plus précisément, il est observé que lors de la traduction anglais-espagnol, l'ambiguïté parallèle présente entre les deux langues permet de ne pas résoudre 50 % des distinctions sémantiques opérées par les lexicographes en anglais. En revanche, dans le cas de langues plus distantes, comme l'anglais et le japonais, 86 % des distinctions sémantiques présentes au niveau monolingue correspondent à des distinctions de traduction, ce qui rend la WSD nécessaire pour la TA. Néanmoins, malgré la divergence des valeurs estimées pour les paires de langues en question concernant la lexicalisation des distinctions sémantiques, d'après Resnik et Yarowsky elles paraissent suffisamment élevées pour justifier une étape de WSD pour la sélection lexicale.

L'influence des langues impliquées dans le processus de traduction sur la performance d'un système de SMT utilisant un module de WSD est également soulignée par Cabezas et Resnik (2005). La sélection lexicale serait plus facile entre langues proches pour un système de SMT basé sur les segments (à l'aide de la table de segments et du modèle de langue)[4] ; en revanche, pour des langues

[3] Ce paramètre influence de manière importante le fonctionnement des méthodes traductionnelles d'analyse sémantique ; ce sujet est analysé dans le paragraphe 2.3.2 du chapitre 2.

[4] Un modèle de SMT nécessite à la fois une méthode qui calcule les probabilités du modèle de langue, une autre qui calcule les probabilités de traduction et une troisième qui cherche la paire de phrases ayant la probabilité maximale. Le **modèle de traduction** d'un système de SMT se base sur les probabilités de traduction des mots des deux langues, tandis que le **modèle de langue** se base sur les probabilités de séquences de mots dans la traduction (Brown *et al*., 1990). La traduction s'effectue par le **décodeur**, qui sélectionne pour une phrase de la LC la phrase de la LS qui a la probabilité maximale.

distantes, où la probabilité de divergences au niveau de l'ordre des mots et de la catégorie grammaticale est plus grande, l'utilisation de techniques de WSD serait plus profitable[5].

Cependant, outre les questions liées à la relation entre les langues impliquées, un autre paramètre ayant un impact sur le besoin d'une étape de WSD dans les systèmes de TA est lié au fonctionnement de ces systèmes.

1.2. Désambiguïsation et sélection lexicale dans les systèmes de SMT

1.2.1. Désambiguïsation indirecte

L'étape de sélection lexicale au sein des systèmes de SMT est étroitement liée à celle de WSD : le but de la WSD est précisément d'aider à la sélection de traductions correctes, d'un point de vue sémantique, pour les mots source.

Les systèmes de SMT actuels effectuent une sorte de **désambiguïsation indirecte**, sans traiter de manière explicite la tâche de désambiguïsation et, par conséquent, sans recours à un module de WSD. En effet, dans les modèles typiques de SMT – comme les modèles IBM (Brown *et al*., 1988, 1990a) –, la sélection entre les candidats de traduction est basée sur le **contexte local** des mots, sans considération d'informations linguistiques plus riches, comme celles exploitées par les systèmes de WSD. Ces informations peuvent être, par exemple, des informations syntaxiques et de collocation locale, des informations relatives à la position des mots dans les textes, des informations morphosyntaxiques relatives aux collocations locales (±1 mot autour du mot à désambiguïser) ou des informations sur les mots environnants à une distance plus grande (Chan *et al*., 2007 ; Carpuat et Wu, 2005a, 2007a, 2007b).

Les architectures de SMT fondées sur les mots réalisent ainsi la désambiguïsation en combinant des **probabilités *a priori*** sur les **candidats de sens** (à partir de critères d'adéquation, modélisés à l'aide de probabilités de traduction des mots) avec des **préférences de cohérence contextuelle** (à partir de critères de fluidité, modélisés par les probabilités du modèle de langue) (Carpuat

[5] Un effet similaire pourrait être observé, d'après Cabezas et Resnik (*ibid*.) dans le cas d'utilisation de corpus hétérogènes vs. corpus homogènes.

et Wu, 2007b). Les architectures de SMT basées sur les segments intègrent en plus, dans les choix de désambiguïsation, des préférences de collocation lexicale.

Néanmoins, malgré l'amélioration significative de la qualité de la traduction apportée par les systèmes de SMT actuels, l'analyse des erreurs montre que des problèmes au niveau de la sélection lexicale persistent (Carpuat et Wu, 2007a). Des modèles de sélection basés sur la WSD, et capables d'incorporer des traits contextuels plus riches que ceux pris en compte par les systèmes typiques de SMT, sont donc souvent considérés comme pouvant aider à la sélection lexicale par de meilleures prédictions, améliorant ainsi la qualité de traduction. Pour une amélioration effective de la précision de WSD au sein des architectures de SMT, il faut que les techniques de WSD utilisées incorporent des hypothèses au moins aussi fortes que celles des modèles de SMT.

1.2.2. Assimilation des tâches de désambiguïsation et de sélection lexicale

1.2.2.1. Désambiguïsation au niveau des mots

La sélection entre les sens d'un mot source dans les systèmes de SMT coïncide, le plus souvent, avec la sélection entre ses traductions dans la LC. La considération implicite de traits contextuels pour la sélection lexicale justifie l'assimilation des tâches de WSD et de sélection lexicale au sein de ces systèmes. Ainsi, le problème de la sélection lexicale est souvent conçu comme un problème de WSD, où les sens correspondent à des mots de la LC. Par conséquent, l'entraînement de **classificateurs de désambiguïsation supervisés** dans un tel cadre peut s'effectuer à partir de textes parallèles alignés au niveau des mots, qui fournissent des paires d'entraînement observables de type ***mot*** / ***sens*** (ce qui correspond dans ce cas à ***mot*** / ***traduction***). Les prédictions concernant les sens des mots à désambiguïser coïncident ainsi avec les prédictions concernant leurs traductions possibles dans la LC.

1.2.2.2. Avantages liés à l'assimilation de la WSD et de la sélection lexicale

L'assimilation des tâches de désambiguïsation et de sélection lexicale présente un certain nombre d'avantages :

(1) Elle permet de **dissocier** entièrement les techniques de WSD d'inventaires de sens prédéfinis (Resnik, 2007). Ces inventaires peuvent contenir des distinctions sémantiques dont la nature n'est pas liée aux données de traduction. L'utilisation d'inventaires de sens extérieurs et non liés aux données bilingues d'entraînement du système de SMT peut donc provoquer une détérioration de la qualité de traduction, comme c'est le cas dans le travail de Carpuat et Wu (2005a).

(2) Il est possible d'éviter le repérage de distinctions de **granularité trop fine**, éventuellement non pertinentes pour la traduction des langues impliquées, en raison des ambiguïtés parallèles (qui peuvent exister). Ce qui importe le plus dans un cadre de traduction est le repérage de relations de traduction entre les mots des deux langues (Vickrey *et al*., 2005)[6]. La considération de la tâche de WSD comme une tâche de sélection lexicale met ainsi l'accent sur les distinctions sémantiques pertinentes pour la traduction.

(3) La conception de la WSD comme sélection lexicale permet de tirer profit de la grande disponibilité de **données étiquetées**, en guise de corpus bilingues alignés au niveau des mots. L'étiquetage de mots source par des mots cible permet d'accroître fortement les données d'entraînement disponibles pour les algorithmes de désambiguïsation supervisés, dans la mesure où il y a davantage de corpus parallèles que de textes sémantiquement étiquetés à la main.

[6] Vickrey *et al.* (2005) soutiennent la non pertinence des nuances sémantiques subtiles pour la sélection lexicale, en raison de la léxicalisation fréquente de sens étroitement liés d'un mot par un seul mot dans une autre langue. Ils proposent donc de se focaliser sur les distinctions sémantiques *considérées par un humain* lors de la traduction. Ce point de vue paraît logique dans une perspective qui touche à la question de l'ambiguïté traductionnelle. Cependant, il est préférable de ne pas affirmer l'inutilité de nuances subtiles de sens pour la sélection lexicale ; les nuances sémantiques fines présentes dans une langue peuvent très bien être lexicalisées dans une autre, ce qui dépend d'un grand nombre de facteurs. De telles nuances peuvent également être véhiculées par les candidats de traduction d'un mot source. Par conséquent, il est recommandé d'aborder la question avec prudence afin d'éviter de telles simplifications.

(4) L'étiquetage à l'aide des traductions permet de lier la WSD (et son évaluation également) à des applications spécifiques (comme la TA et la recherche d'information multilingue). Ce lien peut résoudre la non-conformité de méthodes monolingues de WSD dans un cadre de traduction et même la non-conformité, dans un tel cadre, de méthodes visant d'autres applications (en raison des besoins divergents concernant le degré de WSD, le type et le niveau des distinctions sémantiques dans des applications différentes).

La considération des traductions des mots source en tant que sens de ces mots, parmi lesquels le module de WSD doit sélectionner, est implémentée dans un grand nombre de travaux sur l'utilisation de méthodes de WSD pour la sélection lexicale (Cabezas et Resnik, 2005 ; Vickrey *et al*., 2005 ; Piperidis *et al*., 2005 ; Carpuat et Wu, 2006 ; Chan *et al*., 2007).

Cette manière de concevoir la WSD la rapproche des tâches multilingues de Senseval (Chklovski *et al*., 2004), où les inventaires sémantiques utilisés représentent des distinctions sémantiques effectuées dans d'autres langues.

1.2.2.3. Inconvénients liés à l'assimilation de la WSD et de la sélection lexicale

Des inconvénients d'une telle conception des sens, qui les fait correspondre aux traductions des mots ambigus, concernent le fait que **chaque EQV** est considéré comme décrivant un sens, étant donné qu'aucune distinction n'est faite entre EQVs sémantiquement apparentés et EQVs non apparentés. Le statut des sens n'est donc pas distingué. Les sens sont considérés comme équivalents, ce qui résulte en leur traitement uniforme.

Hormis les problèmes qu'une telle conception du sens soulève au niveau théorique, elle a également un impact négatif au niveau pratique : la sélection entre les différents sens obéit au même type de contraintes et aucune flexibilité n'est permise à ce niveau. En outre, le fait de traiter les sens comme ayant le même statut ne permet pas de prendre en compte l'importance des éventuelles erreurs lors des processus de WSD et de sélection lexicale.

D'après la proposition de Resnik et Yarowsky (1997), la pénalisation des erreurs produites pendant la WSD devrait dépendre de la granularité des distinctions sémantiques concernées. Ainsi, la classification erronée entre sens proches devrait être moins pénalisée que la classification erronée entre sens distants (par exemple, des distinctions entre homographes). Cette idée est liée à l'impact des erreurs de WSD sur la compréhension (pris en compte par la distance fonctionnelle communicative), dans le sens où des erreurs résultant en des malentendus devraient être davantage pénalisées que des erreurs entre distinctions fines.

Dans le cadre de la TA, selon Resnik et Yarowsky, seules les distinctions sémantiques lexicalisées différemment dans la LC devraient être pénalisées[7]. La nécessité de considérer l'impact des erreurs sur la compréhension est également soulignée par Marrafa et Ribeiro (2001), dans un effort de définition des **erreurs lexicales** dans la traduction. Nous reviendrons sur la question de la pénalisation différenciée des erreurs de sélection lexicale dans le chapitre traitant de l'évaluation (chapitre 10).

1.2.2.4. Désambiguïsation au niveau des segments

La focalisation sur les mots simples en tant que cibles pour la désambiguïsation s'explique par le contenu des inventaires sémantiques, aussi bien que par la considération de l'espace comme délimiteur des mots dans les langues européennes (Carpuat et Wu, 2007b). Néanmoins, dans la plupart des systèmes de SMT actuels, l'unité de base de la sélection lexicale est le segment et non le mot (Och et Ney, 2004 ; Koehn, 2004 ; Chiang, 2005). Si les sens des unités source sont supposés correspondre à leurs traductions possibles dans la LC, les « sens » d'un segment source (parmi lesquels la sélection sera effectuée) dans le cas d'un système de SMT basé sur les segments, pourraient être définis comme les segments qui lui correspondent dans la LC.

[7] La distance communicative, dans ce cas, pourrait se fonder sur le pourcentage pondéré de l'ensemble des langues lexicalisant différemment les sens concernés.

Afin que les systèmes de SMT puissent tirer profit des avantages des modèles de WSD[8], une **redéfinition de la WSD** pour la SMT doit être opérée pour l'adapter le plus possible à la tâche à laquelle les systèmes de SMT sont confrontés (Chan *et al*., 2007 ; Carpuat et Wu, 2007a ; 2007b): il s'agit de s'éloigner de la définition simpliste des cibles de désambiguïsation en tant que mots simples, et de s'orienter vers la **désambiguïsation de segments** (*phrase sense disambiguation*). Selon Carpuat et Wu (*ibid*.), cette approche de désambiguïsation, qui généralise la désambiguïsation lexicale à des cibles composées de plusieurs mots (*multi-word targets*), permet d'incorporer dans le modèle de désambiguïsation les hypothèses de base responsables de la réussite des approches de SMT basées sur les segments.

De la même façon que lors de la définition des sens en tant que traductions au niveau des mots, dans le cas de l'utilisation de segments, les sens correspondent aux candidats de traduction des segments rencontrés pendant l'entraînement du système de SMT. Malgré ces adaptations, la tâche de désambiguïsation est toujours considérée comme une véritable tâche de WSD, exploitant des traits contextuels riches qui permettent d'intégrer des probabilités contextuelles relatives aux traductions.

Etant donné les capacités de désambiguïsation des systèmes de SMT, il a été tenté d'estimer précisément leur performance dans une véritable tâche de WSD. Des conclusions intéressantes en ont été tirées, qui seront présentées dans le paragraphe suivant.

1.2.3. Performance des systèmes de SMT en matière de WSD

Carpuat et Wu (2005b) procèdent à une évaluation des capacités effectives des systèmes de SMT en matière de WSD, en comparant leur performance à celle de vrais modèles de WSD, et ce, dans le cadre d'une tâche de WSD précise. L'objectif de cette comparaison empirique est de tester l'hypothèse selon laquelle

[8] Ces avantages consistent en une gamme plus riche de traits, utilisés par les modèles de WSD pour la sélection de sens, et en leur plus grande sensibilité au contexte.

les systèmes SMT sont capables de désambiguïser des mots et que leur capacité de WSD égale celle des modèles de WSD.

L'évaluation en question a été effectuée dans un cadre Senseval, permettant une comparaison directe de la performance des systèmes en matière de WSD, sur un ensemble commun de données. Le système de TA chinois-anglais utilisé est un système de type IBM et le système de WSD est basé sur le modèle ayant la meilleure performance dans la tâche d'échantillon lexical (*lexical sample*) pour le chinois de Senseval-3 (Carpuat *et al.*, 2004)[9]. Les métriques utilisées pour l'évaluation du modèle de SMT sur cette tâche de WSD sont les métriques standard employées pour estimer la précision de la WSD.

Les résultats de cette étude démontrent que la précision de WSD des modèles SMT actuels est significativement **inférieure** à celle des modèles de WSD considérés. Ce qui amène à la conclusion que les modèles actuels de SMT ont des limites en matière de désambiguïsation et que la SMT pourrait bénéficier des meilleures prédictions opérées par les modèles de WSD.

Un ensemble de travaux ont été menés visant à fournir une réponse à la question de la nécessité de passer par une étape de WSD dans les systèmes de SMT. Les conclusions qui en sont tirées se basent sur l'amélioration ou la détérioration des résultats obtenus en matière de qualité de traduction, estimées à l'aide des métriques d'évaluation de la TA. Avant de procéder à une présentation de ces travaux, nous souhaitons analyser le fonctionnement de ces métriques, pour mieux comprendre leur impact sur les résultats.

1.3. Amélioration *vs.* détérioration de la traduction : le rôle des métriques d'évaluation

1.3.1. Qu'est-ce qui est évalué ?

Une estimation humaine de la qualité des traductions produites par un système automatique requiert beaucoup de temps et n'est pas toujours possible.

[9] La tâche d'échantillon lexical de Senseval-3 pour le chinois inclut 20 mots, pour chacun desquels plusieurs sens sont définis à l'aide de la base de connaissances HowNet.

Le recours à une telle évaluation n'est pas évident lors, par exemple, du développement de systèmes de TA, où l'impact de modifications plus ou moins grandes sur le résultat de la traduction doit être estimé au fur et à mesure du développement. Les métriques automatiques proposées pour l'évaluation de la TA tentent de remédier à cet état et d'effectuer une évaluation qui soit la plus proche possible d'une évaluation humaine (White *et al.*, 1994 ; Coughlin, 2003) tout en étant la plus objective possible.

Les métriques automatiques estiment la qualité de la traduction fournie par un système de SMT en calculant sa proximité à une **traduction de référence**. Cette comparaison nécessite donc qu'un corpus de référence contenant des traductions de haute qualité (faites par des humains) soit disponible. Plus une traduction automatique est proche d'une traduction humaine professionnelle, plus elle est considérée comme optimale.

La métrique d'évaluation le plus souvent adoptée est **BLEU** (Papineni *et al*. 2002). Cette métrique estime la proximité d'une traduction candidate à une traduction de référence par la comparaison des *n*-grammes contenus dans les deux traductions, indépendamment de leur position. BLEU repose sur la mesure de **précision** : le nombre de mots de la traduction candidate trouvés dans la traduction de référence est divisé par le nombre total de mots de la traduction candidate[10]. La précision est calculée séparément pour les *n*-grammes de différents ordres et puis les précisions calculées sont combinées en trouvant leur moyenne géométrique pondérée. Une traduction utilisant les mêmes mots (unigrammes) que la traduction de référence est considérée comme satisfaisante du point de vue de l'**adéquation**, tandis que l'analyse de *n*-grammes plus longs correspond à une estimation de la **fluidité** de la traduction[11]. La qualité de la

[10] Pour remédier au problème de sur-génération des systèmes de SMT, la **précision modifiée** de *n*-grammes est proposée, qui prend en compte le nombre maximal de fois qu'un mot peut apparaître dans une traduction de référence.

[11] L'unité de base de l'évaluation est la phrase. Les correspondances de *n*-grammes sont d'abord calculées phrase par phrase ; ensuite, les calculs pour chaque phrase candidate sont additionnés et divisés par le nombre de *n*-grammes candidats dans le corpus de test, afin de calculer un score de précision modifié pour la totalité du corpus. Les précisions modifiées des *n*-grammes sont à la fin combinées. La précision se détériore de manière exponentielle par rapport à *n*, détérioration qui est prise en considération par BLEU (la précision modifiée d'unigrammes est plus grande que la précision modifiée de bigrammes, qui est plus grande que la précision modifiée de trigrammes).

traduction candidate est par conséquent directement proportionnelle au nombre de correspondances à la traduction de référence.

Parmi les inconvénients que comporte cette analyse, on constate l'attribution d'un score élevé à des phrases où le changement de la position des *n*-grammes s'accompagne d'un changement important de sens, ainsi que la pénalisation d'une bonne traduction, lors de l'utilisation de *n*-grammes longs, si l'ordre des *n*-grammes diffère de celui de la référence. Ces effets sont supposés être éliminés par l'utilisation d'un grand nombre de références, qui rend possible une meilleure évaluation en raison du grand éventail de traductions acceptables possibles (Thompson, 1991). Les questions de **variation lexicale**, de **synonymie** et de **paraphrase** sont également supposées être prises en compte par l'utilisation de références multiples (Papineni *et al*., 2002 ; Callison-Burch *et al*., 2006b). Cependant, ce besoin constitue un autre inconvénient de BLEU, étant donné la difficulté de trouver un grand nombre de textes de référence[12].

En outre, selon Callison-Burch *et al*. (*ibid*.) l'utilisation de références multiples aggrave le problème de BLEU concernant les variations autorisées dans la traduction. En effet BLEU, par souci de permettre des variations, ne pose pas assez de contraintes sur l'ordre des *n*-grammes. Cette absence de contraintes rend possible un degré de variation trop élevé lors de l'utilisation de références multiples, beaucoup plus grand que celui considéré comme acceptable dans la traduction. Ainsi, BLEU attribue le même score à des traductions très différentes, n'étant pas toutes aussi pertinentes d'un point de vue sémantique ou syntaxique. Un score BLEU élevé n'est ainsi pas toujours indicatif d'une vraie amélioration de la qualité de traduction.

L'inconvénient supplémentaire de BLEU est de n'exploiter que la notion de précision, sans prendre en compte celle de **rappel** de manière directe. Si tel était le cas, le rappel correspondrait à la proportion des *n*-grammes pour lesquels des correspondances sont trouvées sur le nombre total de *n*-grammes dans la traduction de référence[13]. Pourtant, le rappel constitue un paramètre important de l'évaluation de la qualité d'une traduction, puisqu'il reflète le degré auquel la traduction couvre le contenu entier de la phrase traduite. La raison pour laquelle

[12] Ainsi, l'évaluation se fonde le plus souvent sur un ou deux textes de référence seulement.

[13] Tandis que la précision prend en compte le nombre total de *n*-grammes dans la traduction candidate.

BLEU n'utilise pas le rappel est que cette notion n'est pas clairement établie lorsque des correspondances avec un ensemble de traductions de référence sont recherchées simultanément. Au lieu du rappel, BLEU utilise la **pénalité de la brièveté,** qui pénalise les traductions lorsqu'elles sont trop courtes. Cependant, cette pénalité est considérée comme inadéquate pour compenser l'absence de rappel (Banerjee et Lavie, 2005).

Avec BLEU, l'estimation ne se fonde donc pas sur la possibilité de l'algorithme de traduction de capter et de traduire le sens, mais sur le score obtenu par rapport aux références. Cette faiblesse vaut également pour les autres métriques d'évaluation. **NIST** (Doddington, 2002) évalue la fluidité de la traduction sur la base de correspondances de *n*-grammes avec la référence, mais diffère de BLEU par la prise en compte de la fréquence des *n*-grammes[14] et par l'insensibilité à la casse. **METEOR** (Lavie *et al*., 2004 ; Banerjee et Lavie, 2005 ; Lavie et Agarwal, 2007) combine le calcul de correspondances d'unigrammes avec des informations de WordNet, ce qui permet de capter des cas où des mots synonymes à la référence sont utilisés dans la traduction ; en outre, il permet l'utilisation de mots tronqués (*stemmed*) qui rend possible le repérage de correspondances lexicales lorsque la structure grammaticale de la traduction diffère. METEOR tient également compte de la question de l'ordre des mots dans la phrase, ce qui explique pourquoi des *n*-grammes longs ne sont pas utilisés[15].

F-measure (Turian *et al*., 2003) se fonde également sur les correspondances entre la traduction et la référence en valorisant les correspondances de segments longs. Des mesures fondées sur la **distance d'édition** de la traduction par rapport à la référence ont également été proposées (Alshawi *et al*., 1998 ; Marrafa et Ribeiro, 2001), qui calculent le nombre d'insertions, de suppressions et de substitutions devant être opérées dans la traduction proposée afin d'obtenir la traduction de référence. Ces méthodes, plus simples, présentent néanmoins des inconvénients, comme la double pénalisation d'un mot placé à une position différente dans les deux traductions comparées, en raison de la nécessité de sa

[14] Des *n*-grammes apparaissant souvent et véhiculant peu d'informations sont considérés par BLEU comme ayant un impact égal sur la précision totale que les *n*-grammes véhiculant des informations riches. La quantité d'informations d'un *n*-gramme est, par contre, estimée par NIST relativement à sa fréquence d'apparition ; une corrélation négative est supposée exister entre les deux.

[15] Une pénalisation sur le changement d'ordre dans la phrase est calculée, qui dépend du nombre des segments dans le texte produit qui doivent être déplacés pour obtenir le texte de référence.

suppression de la position occupée et de son insertion dans une position différente (Bangalore *et al*., 2000).

1.3.2. Quantification de l'impact de la WSD sur le résultat de la SMT

On constate donc qu'en général les métriques d'évaluation de la TA sont très **strictes**, nécessitant des correspondances exactes entre les occurrences présentes dans la traduction proposée par le système et la traduction de référence. Par conséquent, lorsqu'elles sont utilisées pour estimer la qualité de la traduction d'un système utilisant une méthode de WSD, elles pénalisent souvent des décisions de traduction correctes du point de vue qualitatif ne correspondant pas à la référence.

Les solutions proposées pour remédier à cette faiblesse des métriques d'évaluation, comme l'utilisation de références multiples (BLEU) ou l'exploitation d'inventaires sémantiques (METEOR), se heurtent néanmoins à d'autres problèmes : dans le premier cas, il s'agit, comme nous l'avons déjà souligné, de la non disponibilité d'un grand nombre de références et des complications au niveau des calculs (Callison-Burch *et al*., 2006b) ; dans le deuxième cas, il s'agit de la restriction de l'application de la métrique aux langues disposant des ressources sémantiques requises (Lavie et Agarwal, 2007).

La difficulté à prendre en compte les améliorations qualitatives effectuées par le module de WSD est soulignée par tous les travaux traitant de l'intégration d'un tel module dans un système SMT. Carpuat et Wu (2005a) remarquent une détérioration du score BLEU lors de l'utilisation de la traduction sélectionnée par le modèle de WSD, même dans les cas où elle est meilleure que celle proposée par le modèle de TA. Ce phénomène s'explique par le souci de **cohérence phrastique** : la traduction sélectionnée par le modèle SMT est plus probable, d'après le modèle de langue, que la prédiction du modèle de WSD ; si ce n'était pas le cas, la prédiction du modèle de WSD serait également sélectionnée par le système de SMT. Les systèmes de SMT traduisant des phrases complètes, et non des mots isolés, effectuent la sélection lexicale d'une manière qui privilégie la cohérence dans la traduction (calculée par le modèle de langue). Les prédictions bénéficient ainsi du contexte phrastique de la LC, ce qui améliore la fluidité de la

traduction et, par conséquent, le score BLEU. L'utilisation de la prédiction d'un modèle de WSD, prédiction qui n'a pas été rencontrée assez fréquemment au sein des segments, provoque une baisse du score BLEU ; tel est le cas même lorsqu'il s'agit d'une prédiction correcte, en raison du fait que le modèle de SMT ne sache pas comment utiliser cette prédiction correctement. Cet effet est appelé **effet du modèle de langue**.

La pénalisation des changements effectués lors de l'exploitation des propositions du module de WSD est également remarquée dans les travaux qui font état d'une amélioration du score BLEU (Cabezas et Resnik, 2005 ; Chan *et al*., 2007). Cette pénalisation est effectivement liée à l'absence de correspondances au niveau des *n*-grammes entre la traduction obtenue et la référence.

D'un point de vue quantitatif, malgré le fort impact négatif des métriques d'évaluation sur les résultats de la SMT utilisant des méthodes de WSD, des aspects intéressants sont mis en évidence par les travaux visant à répondre à la nécessité d'une telle étape dans les systèmes de SMT. Ces aspects concernent le cadre au sein duquel les expériences sont menées, la façon d'intégrer des modules de WSD dans les systèmes de SMT et les hypothèses adoptées.

1.4. Impact négatif de la WSD sur la qualité du résultat de la TA

1.4.1. Détérioration de la qualité de traduction par intégration d'un module de WSD

Le travail de Carpuat et Wu (2005a) ayant démontré l'impact négatif de la WSD sur le résultat d'un système de SMT, il a constitué une référence pour les travaux ultérieurs sur la question (qui ont pris le contre-pied de cette position). En effet, ce travail nie l'hypothèse fort répandue concernant le rôle important de la WSD dans les systèmes de TA et met en cause son statut en tant qu'étape nécessaire ou, au moins, bénéfique, au sein de ces systèmes. Cette conclusion est tirée d'une comparaison de la sortie d'un système de SMT avec et sans exploitation des résultats d'un processus de WSD.

L'étude de Carpuat et Wu concerne, plus précisément, une tâche de traduction chinois-anglais, qui implique l'utilisation d'un système de WSD *état de l'art* et d'un modèle typique de SMT basé sur les mots (de type IBM). Le système de WSD est un système supervisé de type SENSEVAL[16], dont l'objectif est de prédire des traductions anglaises des mots ambigus chinois en contexte. En absence de données d'apprentissage chinoises annotées par des sens anglais, le système est entraîné sur les données de la tâche d'échantillon lexical (*lexical sample*) de Senseval-3 pour le chinois, qui concerne 20 mots ambigus, dont les sens sont définis à l'aide de la base de connaissances HowNet (Dong, 1998). La prédiction de traductions est rendue possible par l'exploitation des mots anglais (*gloss*) décrivant chacun des sens dans HowNet.

Les prédictions du système de WSD sont intégrées dans le modèle de SMT de deux manières : (a) pendant le **décodage** et (b) pendant une étape de **post-traitement**. Dans le premier cas, les prédictions effectuées par le modèle de traduction ne sont pas considérées et le décodeur est **forcé** de choisir la meilleure traduction à partir des mots anglais illustrant le sens sélectionné par le système de WSD. Dans le deuxième cas, la traduction du mot ambigu sélectionnée par le modèle de SMT est **remplacée** dans le segment de sortie par la prédiction du système de WSD[17].

La qualité des traductions fournies par le système avec et sans WSD est ensuite évaluée à l'aide de la métrique BLEU (Papineni *et al*., 2002). La méthode qui force le décodeur à sélectionner une traduction parmi celles proposées par le système de WSD détériore la qualité de la traduction, à cause du modèle de langue[18]. Nous rappelons que l'effet du modèle de langue souligne une des faiblesses du score BLEU, à savoir qu'il pénalise davantage les choix ayant un effet négatif sur la **cohérence** du texte traduit que ceux qui ne sont pas les meilleurs du point de vue sémantique, sacrifiant ainsi la **pertinence** à la **fluidité** (cf. §1.3.2.).

[16] Il s'agit du système qui a eu la meilleure performance dans la tâche « Senseval-3 Chinese lexical sample » (Carpuat *et al*., 2004).
[17] Lorsque plusieurs traductions sont proposées, une traduction est choisie de manière aléatoire.
[18] La qualité du résultat baisse davantage dans le cas où les prédictions du modèle de WSD sont enrichies par des mots anglais ayant une probabilité élevée d'après le modèle de traduction.

L'effet du modèle de langue est considéré être éliminé dans l'approche permettant d'utiliser les prédictions de WSD pendant le post-traitement, où le décodage est effectué avant de connaître les prédictions du modèle de WSD. Cette approche donne des meilleurs résultats que celle de décodage, ce qui démontre effectivement l'incapacité du modèle de langue d'utiliser correctement les prédictions de WSD. Cependant, même dans ce cas, la performance du modèle SMT sans WSD est meilleure.

1.4.2. Discussion autour de l'impact négatif de la WSD sur le résultat de la Traduction Automatique

Les résultats obtenus par Carpuat et Wu (2005a) ont déclenché un débat très intéressant autour de la nécessité de WSD pour la TA. Dans les travaux ultérieurs sur la question, les auteurs prennent tous le soin de se positionner par rapport aux conclusions tirées de ces expériences concernant l'impact négatif de la WSD sur la TA, en clarifiant les facteurs qui ont provoqué ces résultats, voire même en critiquant le cadre dans lequel les expériences ont été menées. Ces critiques nous semblent très intéressantes dans la mesure où elles aident, d'une part, à comprendre les résultats « surprenants » obtenus par Carpuat et Wu et où, d'autre part, elles montrent ce qui doit être évité dans une telle étude. C'est pour cette raison que nous souhaitons en faire une synthèse, avant de procéder à la présentation des méthodes qui contredisent cette conclusion.

Les principales raisons ayant contribué à la détérioration de la qualité de traduction observée par Carpuat et Wu (2005a), lors de l'utilisation d'un module de WSD, sont les suivantes :

(a) Le petit volume de données d'entraînement du système de WSD utilisé (environ 37 instances d'entraînement pour chaque mot) (Cabezas et Resnik, 2005).

(b) L'absence de lien entre les données d'entraînement du système de WSD et les données bilingues d'entraînement du système de SMT (Cabezas et Resnik, *ibid*.).

(c) L'absence de lien entre l'inventaire de sens utilisé (HowNet) et les données d'entraînement du système de SMT (Cabezas et Resnik, *ibid.* ; Chan *et al.*, 2007). Le sens d'un mot est d'abord choisi à partir de HowNet par le module de WSD et la prédiction de traduction se base ensuite sur l'ensemble de mots anglais associés à ce sens, bien que les auteurs admettent que l'idéal serait de considérer directement les traductions anglaises en tant que sens (Chan *et al.*, *ibid.*).

(d) L'intégration des prédictions du système de WSD à l'aide de contraintes « dures » (Cabezas et Resnik, *ibid.*, Vickrey *et al.*, 2005) et la non intégration du modèle de WSD et de ses prédictions dans le modèle de traduction (Chan *et al.*, *ibid.*) ; au contraire les prédictions sont exploitées soit en forçant le décodeur à effectuer un choix, soit en remplaçant les traductions choisies par le système de SMT lors d'une étape de post-traitement. Ce problème d'intégration des résultats d'un système de WSD dans les traductions caractérise, d'après Vickrey *et al.*, de manière générale les décodeurs actuels, dans la mesure où ils n'offrent pas la possibilité d'intégrer naturellement dans la traduction les résultats d'un module de sélection lexicale.

(e) L'utilisation du score BLEU pour l'évaluation, dont les faiblesses ont été déjà présentées (cf. §1.3.1.). Callison-Burch *et al.* (2006b) proposent une re-évaluation manuelle du travail de Carpuat et Wu.

Les travaux plus récents sur l'impact de la WSD dans un système de SMT rapportent tous une **amélioration** des résultats de traduction. Nous exposons ces travaux dans le paragraphe suivant, en soulignant leurs points communs et leurs divergences.

1.5. Impact positif de la WSD sur la qualité du résultat de la TA

1.5.1. Points communs et divergences des méthodes

Les travaux démontrant l'impact bénéfique de l'intégration d'un module de WSD dans un système de SMT présentent certaines caractéristiques communes. Tout d'abord, les sens candidats à la WSD ne sont pas issus d'inventaires de sens définis *a priori*, mais de corpus bilingues. Plus concrètement, dans ces méthodes, les sens des mots (ou des segments) sont désignés par leurs traductions au sein de corpus parallèles alignés.

En outre, l'entraînement des systèmes de WSD utilisés est effectué sur les mêmes corpus que l'entraînement du système de SMT, et non sur des données d'apprentissage manuellement annotées.

Un autre point commun des méthodes en question se situe au niveau de l'intégration des prédictions de WSD dans les systèmes de SMT, qui est faite de manière dynamique. Ainsi, ces prédictions constituent des alternatives que le décodeur prend en compte mais sans être contraint de sélectionner la traduction proposée par le système de WSD.

Certaines divergences existent néanmoins entre ces méthodes, quant à la complexité des problèmes de WSD et de traduction traités. Certaines ne visent la distinction que d'un petit nombre de sens, tandis que d'autres proposent des solutions à des problèmes simplifiés de traduction. De plus, des divergences apparaissent au niveau des systèmes de SMT utilisés dans le cadre des expériences, provoquant ainsi un impact sur la définition des cibles de désambiguïsation. Une présentation plus détaillée de ces méthodes et de leurs résultats est donnée dans les paragraphes qui suivent.

1.5.2. Amélioration de la traduction par résolution d'un problème simplifié de WSD

La première étude sur la question de l'impact de la WSD sur le résultat de la TA a été menée par Brown *et al*. (1991b). Cette étude montre l'effet bénéfique de l'intégration d'un module de WSD dans un système de SMT. Néanmoins, leur

méthode de WSD est très limitée, en ce qu'elle ne permet la distinction qu'entre **deux sens** d'un mot source, traduits différemment dans la LC. Cette méthode permet la prise en compte d'informations plus riches que le modèle de langue du système de SMT. Ce modèle parvient parfois à corriger la traduction d'un mot proposée par le modèle de traduction, en prenant en compte le contexte du mot dans la LC (à l'aide de trigrammes). Cependant, comme il ne capte que des phénomènes locaux, si l'élément du contexte permettant l'identification du sens d'un mot ambigu se situe hors de sa portée (à une distance supérieure à 3 mots), la désambiguïsation de ce mot n'est alors pas possible, ni l'éventuelle correction de la traduction.

La méthode de WSD de Brown *et al*. (*ibid*.) se fonde, quant à elle, sur l'idée que lorsque le contexte nécessaire pour la WSD se trouve hors de la portée du modèle de langue, il peut être **localement encodé** et ce afin d'éviter une traduction erronée du mot. Cette méthode permet l'**étiquetage** des mots source de manière à élucider leurs traductions dans la LC. Ainsi, étant donné une connexion entre deux mots de deux langues, la méthode de WSD permet l'étiquetage du mot source par un sens (« premier » ou « deuxième » sens), à l'aide de questions binaires portant sur les éléments de son contexte[19]. Une question est construite pour chaque trait contextuel pouvant apporter des informations à cet égard (appelé **informateur**)[20] ; la question fournissant l'IM la plus élevée pour la traduction du mot indique le meilleur informateur. Les traductions possibles du mot ambigu sont ainsi réparties en deux ensembles, chacun correspondant aux traductions les plus probables pour un des sens identifiés.

En intégrant cette méthode de WSD dans leur système de SMT, Brown *et al*. constatent une **nette diminution** du taux d'erreur du système. La qualité de traduction, avec et sans utilisation de la méthode de WSD, a été manuellement

[19] L'information mutuelle (IM) entre les mots des deux langues est calculée sur la base de la probabilité de leur connexion. Leur étiquetage par des sens se fait de manière à ce que l'IM entre les membres de la connexion augmente.

[20] L'informateur peut être un mot à gauche ou à droite, le premier nom à gauche ou à droite, le premier verbe à gauche ou à droite, le temps du mot (s'il s'agit d'un verbe) ou du premier verbe à gauche. Les informateurs sont dépendants de la langue. Ceux décrits ci-dessus sont valables dans le cas d'un mot français ; pour un mot anglais, les questions concernent seulement le premier et le deuxième mot à gauche.

évaluée. Néanmoins, même si les résultats sont prometteurs, cette méthode demeure très limitée, puisqu'elle ne permet d'attribuer que deux sens à un mot.

1.5.3. Correspondances entre sens et traductions pour la WSD dans un cadre de SMT

1.5.3.1. Présentation des méthodes

Des travaux plus récents contredisent les conclusions de Carpuat et Wu (2005a), en présentant des résultats qui démontrent l'effet positif de la WSD sur les résultats des systèmes de SMT. Dans le cadre de ces travaux, la WSD n'est pas restreinte à un nombre précis de sens, mais les sens des mots sont représentés par leurs traductions dans un bitexte aligné au niveau des mots.

Cabezas et Resnik (2005), par exemple, étiquettent les instances des mots ambigus source dans le corpus d'entraînement par leurs traductions dans la LC. Le système de WSD qu'ils emploient, qui est un système d'apprentissage supervisé, traite ensuite chaque mot ambigu comme un problème de classification indépendant. Des traits contextuels pondérés sont attribués à chaque instance annotée du mot, traits qui sont ensuite exploités par un classificateur pour attribuer à l'instance sa catégorie[21]. Les résultats du système de WSD sont intégrés dans le système de SMT en tant qu'alternatives (accompagnées de probabilités)[22], que le décodeur considère en même temps que les alternatives proposées par le modèle de traduction ; la sélection finale est effectuée par le modèle de langue.

Chan *et al*. (2007) intègrent dans leur système de SMT (Chiang, 2005) un système de WSD capable d'établir des prédictions aux niveaux des mots et des segments. Le système de SMT utilisé est un système hiérarchique basé sur les segments (Hiero), reposant sur une grammaire hors contexte synchrone pondérée. Les règles de la grammaire sont automatiquement extraites à partir

[21] Les traits utilisés sont à la fois des traits locaux (± 3 mots) et des traits du contexte lointain (mots de la phrase, de la phrase précédente et de la phrase suivante).

[22] Le système de SMT utilisé donne la possibilité de marquage XML dans le segment source permettant d'indiquer des possibilités de traduction d'une séquence de mots. Il s'agit du système Pharaoh (Koehn, 2004), qui est un système SMT basé sur les segments.

d'un bitexte aligné au niveau des mots[23] et sont ensuite utilisées pour traduire une phrase source, en repérant sa dérivation la plus probable. Les prédictions du système de WSD sont utilisées pour aider Hiero à obtenir une meilleure dérivation lors de la traduction d'une phrase source (en chinois).

Le contexte d'apprentissage pendant l'extraction d'une règle de grammaire pour un segment source est constitué de la phrase qui le contient et des phrases précédente et suivante[24]. Les sens du segment chinois consistent en les segments anglais qui le traduisent dans le bitexte d'entraînement. La sortie du classificateur de WSD consiste, quant à elle, en l'ensemble des traductions anglaises du segment source associées à des probabilités contextuelles[25].

Lorsqu'une règle grammaticale est traitée pendant le décodage, si certains symboles terminaux de la LC sont également proposés comme traductions par le système de WSD, deux traits sont alors calculés : un trait qui fournit la probabilité contextuelle du classificateur de WSD pour la traduction d'un mot source par un mot de la LC, et un autre trait qui récompense les règles utilisant les traductions proposées par le module de WSD. Si la règle est finalement retenue comme partie de la meilleure dérivation de la phrase chinoise, tous les mots du côté anglais de la règle apparaissent dans la traduction. Le réglage des poids des traits et le décodage placent le système de WSD au même niveau que les autres composantes du modèle.

Dans leurs travaux plus récents, Carpuat *et al*. (2006) et Carpuat et Wu (2007a, 2007b) se positionnent également du côté de ceux qui défendent l'impact positif de la WSD sur le résultat de la SMT. Le système de SMT qu'ils utilisent (Pharaoh) diffère de celui utilisé dans leur travail antérieur (Carpuat et Wu, 2005a). En outre, l'entraînement du système de WSD ne se fait pas sur des données d'apprentissage manuellement annotées mais sur les corpus utilisés par

[23] Une règle de grammaire hors contexte synchrone consiste en une portion chinoise et une portion anglaise correspondante, où chaque portion est une séquence de mots et de symboles non terminaux.

[24] Les sources de connaissance utilisées par le classificateur de WSD sont les collocations locales d'une instance du mot ambigu (± 1, - 1, +1), les parties du discours (de ce mot et des mots ±1) et des unigrammes du contexte (même dans une phrase différente du mot ambigu).

[25] Après la WSD, chaque mot d'une phrase chinoise peut avoir jusqu'à trois ensembles de traductions associés : un ensemble pour le mot tout seul ; un ensemble pour ce mot et le mot qui le précède, traités comme une unité ; un ensemble pour ce mot et le mot qui le suit, traités comme une unité.

le système de SMT. Ainsi, les sens des mots correspondent à leurs candidats de traduction rencontrés lors de l'entraînement du système de SMT dans Carpuat *et al*. (2006), tandis que dans Carpuat et Wu (2007a, 2007b) l'entraînement du module de désambiguïsation et les prédictions portent directement sur les segments (cf. § 1.2.2.4.).

Dans le premier cas, pour chaque mot d'une phrase d'entrée du système de SMT, qui a été rencontré dans les données d'entraînement, un modèle de WSD et une distribution dépendante du contexte concernant ses candidats de traduction sont disponibles. Cette distribution contribue à enrichir le lexique de base[26]. Lors du décodage, les hypothèses de traduction construites à partir de ces entrées supplémentaires du lexique entrent en compétition avec celles construites au sein du lexique de segments de base (Carpuat *et al*., 2006). Dans le cas de la désambiguïsation des segments, les prédictions sont définies pour chaque entrée du lexique du système de SMT et elles peuvent être conçues comme des traits additionnels au sein de ce lexique. Contrairement aux probabilités de traduction de SMT typiques, ces prédictions sont sensibles au contexte et nécessitent d'être mises à jour pour chaque nouvelle phrase. Ainsi, au lieu d'utiliser un lexique de traduction de segments statique, l'intégration des prédictions effectuée requiert une mise à jour dynamique du lexique de traduction de segments pour chaque phrase lors du décodage (Carpuat et Wu, 2007a, 2007b).

1.5.3.2. Résultats indiquant une amélioration de la qualité de traduction

Les travaux présentés ci-dessus démontrent tous une amélioration de la qualité des résultats des systèmes de SMT par intégration d'un système de WSD.

L'amélioration du score BLEU obtenue par Cabezas et Resnik (2005) lors de la considération des prédictions de WSD par le système de SMT, et ce, par rapport au cas où le système ne prend pas en compte ces prédictions (qui sert de méthode de base) est faible[27]. Néanmoins, tout en soulignant l'impact du caractère strict de la métrique utilisée sur les résultats, les auteurs adoptent une

[26] A l'instar de Cabezas et Resnik (2005), Carpuat et Wu se servent du schème de marquage XML fourni par Pharaoh pour spécifier les candidats de traduction et leurs probabilités.

[27] La signification statistique de l'amélioration rapportée n'est pas déterminée, ce qui paraît peu concluant (Chan *et al*., 2007).

attitude optimiste quant à l'amélioration de la traduction par prise en compte de la WSD. Cette attitude ce justifie par la petite amélioration quantitative observée, couplée aux résultats d'une évaluation qualitative.

Certaines améliorations proposées concernent la prise en compte de la catégorie grammaticale du mot ambigu[28] et d'un contexte plus large, fournissant des informations sur le sujet traité. Autre possibilité, l'application sélective de la WSD. En cas de distributions de sens asymétriques, par exemple, l'utilisation du sens prédominant ou de celui proposé par le décodeur pourrait suffire. Des techniques d'évaluation de la fiabilité pourraient permettre d'exploiter la solution proposée par le module de WSD seulement si sa probabilité est supérieure à celle de la proposition du système de SMT.

L'analyse des résultats obtenus par Chan *et al*. (2007) montre que la WSD améliore la sortie de leur système de SMT de deux manières : premièrement, la traduction est plus **complète** que celle du fonctionnement basique du système, car davantage de mots anglais appropriés sont proposés ; deuxièmement, la WSD permet de **corriger** des traductions incorrectes. Malgré la difficulté qu'a la métrique d'évaluation utilisée à prendre en compte les propositions du module de WSD par la, Chan *et al*. (*ibid*.) signalent une amélioration significative au niveau statistique. Il est par ailleurs important de souligner que le système Hiero de base donne de meilleurs résultats que Pharaoh (Koehn, 2004), utilisé dans la plupart des autres expériences.

Enfin, Carpuat *et al*. (2006) et Carpuat et Wu (2007a, 2007b) rapportent aussi des améliorations de la performance de leur système de SMT, essentiellement dans les cas où la désambiguïsation porte directement sur les segments. L'intégration du système de désambiguïsation de segments dans le système de SMT produit en effet des améliorations, certes faibles mais cohérentes du point de vue de la qualité de traduction, améliorations constatées dans un ensemble de tests fondés sur huit métriques d'évaluation différentes.

[28] La version du système présentée dans Cabezas et Resnik (*ibid*.) ne prend en compte la catégorie grammaticale du mot à traduire que dans le cas où les traits locaux fournissent davantage de fiabilité pour des traductions dans une catégorie que dans une autre.

1.5.4. Impact positif de la WSD sur des problèmes simplifiés de traduction

Vickrey *et al*. (2005) évaluent l'impact de la WSD sur deux problèmes simplifiés de traduction : la **traduction des mots** et le **remplissage de vides** (*blank-filling*). Dans cette étude, la tâche de WSD est clairement reformulée comme une tâche de traduction de mots. L'objectif de cette étude est de montrer que des techniques d'apprentissage automatique peuvent être utilisées de manière efficace pour la traduction de mots.

Ici aussi, les choix de traduction pour un mot sont définis comme l'ensemble de mots qui lui sont alignés au sein du bitexte d'entraînement. Les traits utilisés par le modèle de traduction concernent la partie du discours du mot ambigu et la présence d'autres mots autour de lui, dans des fenêtres textuelles de tailles variées. Les résultats du système de WSD de Vickrey *et al*. peuvent être intégrés dans les traductions de manière « souple » : plutôt que de forcer l'adoption du premier choix du module de sélection lexicale, le module de traduction de mots peut collaborer avec le modèle de langue et le modèle d'alignement afin de produire la traduction la plus pertinente.

Les résultats obtenus montrent une amélioration de la précision de leur modèle sur les deux tâches simplifiées de TA qui sont envisagées. Néanmoins, leur module de traduction de mots n'a pas été intégré dans un système de TA complet et leur évaluation ne concerne pas la qualité de la traduction de l'ensemble de la phrase source (Chan *et al*., 2007).

Après avoir exposé les conclusions issues de différentes études portant sur les améliorations que peut représenter l'intégration d'un module de WSD dans un système de TA, nous allons désormais présenter une méthode de WSD basée sur le clustering sémantique effectué par notre méthode d'acquisition de sens.

Les sens utilisés sont repérés à partir des données du bitexte d'apprentissage et sont donc aptes à être exploités dans une tâche de WSD pour la traduction. Cette méthode pourrait donc servir à créer un module de WSD utilisable dans un système de TA.

2. Désambiguïsation basée sur le clustering sémantique

2.1. Considération de relations complexes entre sens et équivalents

Les méthodes de WSD utilisées dans les systèmes de SMT décrites jusqu'ici s'appuient sur des informations contextuelles riches relatives à un mot ambigu (ou à un segment) afin de sélectionner le sens d'une nouvelle instance du mot (ou du segment). Les informations exploitées par ces méthodes sont des informations du contexte proche et lointain du mot source, parfois accompagnées d'informations morphosyntaxiques relatives aux traits contextuels retenus.

Dans ce cadre de désambiguïsation, les sens d'un mot ambigu sont identifiés à l'aide de ses équivalents de traduction, chaque équivalent étant considéré comme correspondant à un sens différent du mot. L'identification du sens d'une instance coïncide donc avec la sélection de sa traduction. Des questions telles que la nature des sens lexicalisés par chacun des équivalents ou leurs relations sémantiques ne se posent pas[29]. En effet, il est vrai que, dans les applications du TAL, la distinction des instances d'un mot ambigu en fonction des sens selon lesquels elles sont utilisées peut s'avérer suffisante et que la nécessité d'avoir recours à des connaissances sur la nature des sens impliqués ne se pose même pas (Ide *et al*., 2002).

Néanmoins, au niveau théorique, la considération de correspondances biunivoques entre sens et équivalents paraît simpliste et peut facilement être mise en question. La complexité des relations pouvant exister entre un mot source et ses équivalents a déjà été montrée dans le paragraphe 2.3. du chapitre 1. De plus, la considération de relations de type **un à un** entre sens et équivalents ne permet pas la prise en compte des relations éventuellement existantes entre équivalents similaires du point de vue sémantique.

Notre méthode de WSD permet, quant à elle, de traiter ces relations en se fondant sur les résultats du clustering effectué lors de l'étape d'acquisition de sens (cf. chapitre 6). Les sens d'un mot source sont également définis ici à l'aide des équivalents du mot ; néanmoins, à la différence des méthodes précitées, la

[29] Sauf dans le travail de Carpuat et Wu (2005a), où les équivalents envisagés sont ceux qui correspondent au sens de HowNet sélectionné pour le mot ambigu.

relation entre sens et équivalents peut être de type **un à plusieurs** ou, inversement, **plusieurs à un**. Le premier type se rencontre dans les cas où le sens source est repéré à l'aide d'un cluster d'EQVs. Dans ce cas, les correspondances de traduction entre les mots des deux langues sont enrichies par des informations se situant sur l'**axe paradigmatique**, voire par des informations à propos de la similarité sémantique entre les équivalents. Cette possibilité permet de proposer des traductions alternatives sémantiquement pertinentes du mot source et apporte une certaine flexibilité au processus de sélection lexicale. Le deuxième type de relation concerne les cas où un EQV est considéré comme pouvant traduire plusieurs sens du mot source. Cette relation permet de prendre en compte le phénomène de l'ambiguïté parallèle, où la sémantique de l'EQV est similaire à celle du mot source.

2.2. Exploitation des résultats de l'acquisition de sens pour la WSD

2.2.1. Désambiguïsation sur la base des clusters de sens

2.2.1.1. Nouvelles instances des mots ambigus

La méthode de WSD proposée s'appuie sur les résultats du processus d'acquisition de sens, qui établit des correspondances entre les mots source et les clusters de leurs EQVs. L'entrée de la méthode est constituée d'une phrase contenant une nouvelle instance du mot ambigu et la sortie concerne le cluster décrivant le sens du mot dans le nouveau contexte, comme décrit dans la figure 1.

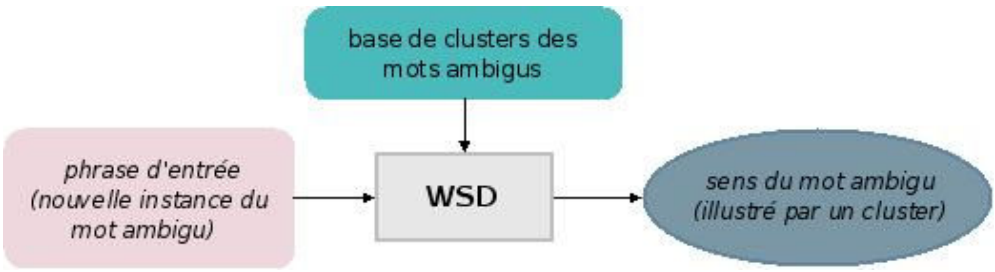


Figure 1. Entrée et sortie de la méthode de désambiguïsation lexicale

Les nouvelles instances proviennent du corpus de test, qui est la partie anglais-grec du corpus parallèle EUROPARL (Koehn, 2003, 2005)[30]. Nous rappelons que des **sous-corpus de test** ont été constitués à partir de ce corpus, correspondant à chaque mot ambigu (cf. §2.2.4, chapitre 4). Chaque sous-corpus a été ensuite filtré en fonction des EQVs du mot source (cf. §2.2.5, chapitre 4). L'EQV traduisant l'instance du mot ambigu dans les unités de traduction retenues constitue la **traduction de référence**. Cette référence servira, par la suite, à l'évaluation de la méthode de WSD (et de la méthode de sélection de traduction, comme nous le montrerons plus loin).

2.2.1.2. Traits des clusters utilisés pour la WSD

Les informations du nouveau contexte exploitées pour la WSD consistent en les cooccurrents du mot ambigu au sein de la phrase d'entrée. Ces informations sont comparées à celles caractérisant les correspondances sémantiques inter-langues, établies pendant l'apprentissage. La correspondance entre le mot ambigu et un cluster peut être définie par trois types d'informations :

a. dans le cas d'un cluster à **1 EQV** : l'ensemble des traits pertinents retenus des contextes de l'EQV

b. dans le cas d'un cluster à **2 EQVs** : les contextes assimilateurs des EQVs (cf. §4.3.2.1, chapitre 7)

c. dans le cas d'un cluster à **plus de 2 EQVs** : l'intersection des contextes assimilateurs des paires des EQVs contenus dans le cluster (cf. figure 2).

[30] Les caractéristiques du corpus et les étapes de prétraitement qu'il a subies sont présentées en détail dans le paragraphe 2 du chapitre 4.

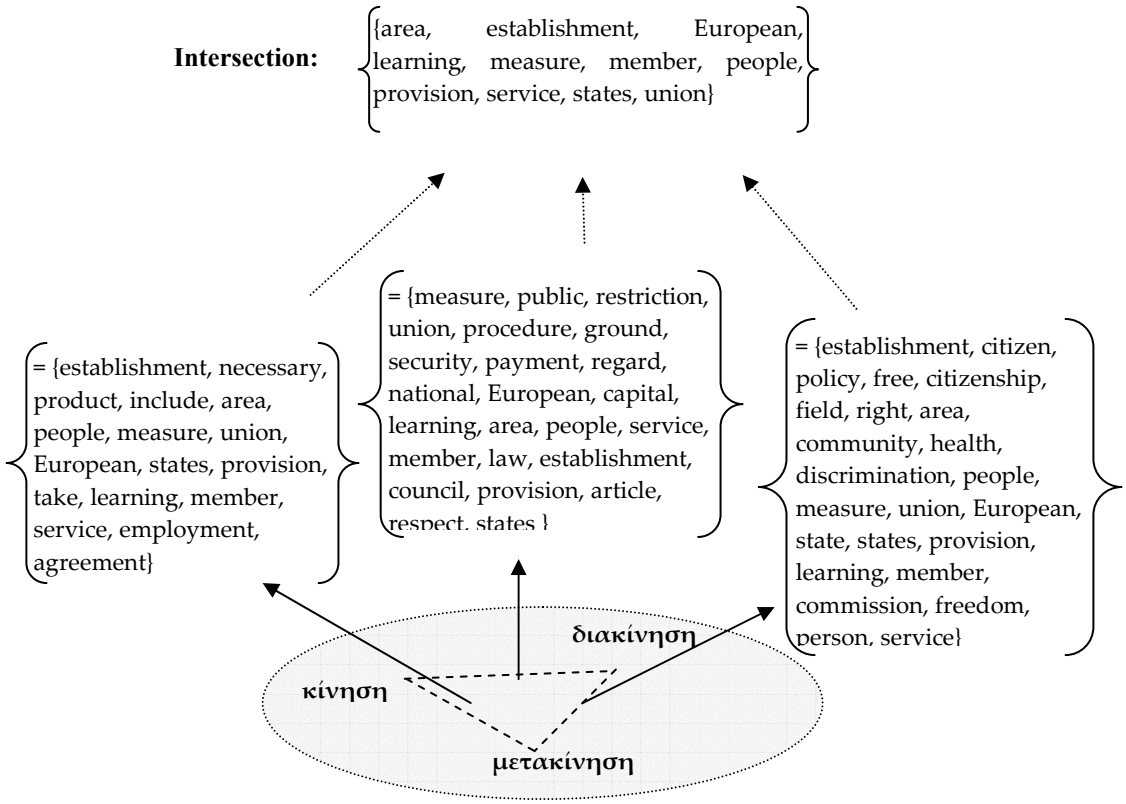


Figure 2. Intersection des contextes assimilateurs des EQVs clustérisés

Les quatre clusters formés pour le mot *movement* sont décrits par les ensembles de traits contextuels suivants :

1. {***διακίνηση, μετακίνηση, κίνηση***} : *area, establishment, European, learning, measure, member, people, provision, service, states, union*
2. {***διακίνηση, κινητικότητα, μετακίνηση***} : *citizen, european, freedom, member, person, states*
3. {***διακίνηση, κίνηση, κυκλοφορία***} : *area, article, capital, council, establishment, European, law, learning, measure, member, national, payment, people, procedure, provision, public, regard, respect, security, service, states, union*
4. {***κίνημα***} : *Olympic, authority, spawn, follow, European, new, world, Rachel, change, certain, teaching, include, Europe, country, vote, book, political, landmark, publication, public, disability, agency, create, ddt, foreign, other, use,*

conference, *special*, *rights*, *patients*, *obligation*, *language*, *spring*, *environmental*, *silent*, *first*

2.2.1.3. Traits des nouveaux contextes

Les traits qui caractérisent les clusters sont les formes de mots de catégories grammaticales précises. Pour que la comparaison de ces traits avec les informations des nouveaux contextes soit possible, ces contextes sont lemmatisés et étiquetés morphosyntaxiquement (cf. §2.2.3, chapitre 4). Les informations retenues du nouveau contexte concernent donc les **formes** des **noms**, des **adjectifs** et des **verbes** qui y apparaissent. Soit la phrase à traiter suivante, qui contient une nouvelle instance du mot ambigu *movement* :

*On the internal market there has been a standstill on many issues, from the free **movement** of persons to the European company statute, to taxation, to the banking and insurance sector.*

L'ensemble des informations de cooccurrence retenues, qui sert à décrire le contexte de cette instance de *movement*, est le suivant :

{*internal* (JJ), *market* (NN), *have* (V), *be* (V), *standstill* (NN), *many* (JJ), *issue* (NN), *free* (JJ), *person* (NN), *European* (JJ), *company* (NN), *statute* (NN), *taxation* (NN), *banking* (NN), *insurance* (NN), *sector* (NN)}

Cet ensemble de traits est comparé aux traits décrivant les clusters afin de choisir celui qui décrit le sens de la nouvelle instance de *movement*.

2.2.1.4. Comparaison entre contexte et clusters

Pour la désambiguïsation, l'ensemble des traits retenus du nouveau contexte est comparé séparément aux traits caractérisant chaque cluster. Cette comparaison sert à trouver l'**intersection pondérée** des deux ensembles de traits.

On définit l'intersection pondérée comme suit : si une communauté de traits existe entre le nouveau contexte et un cluster, une **association** caractérisée comme **pondérée** est alors établie entre eux. Cette association est déterminée par les traits de leur intersection et par le poids attribué à l'association par la fonction *renvoie_intersection_pondérée*. Le processus de pondération est décrit en détail

dans le paragraphe suivant. Si **une seule association** est établie entre le nouveau contexte et les clusters, ce cluster est retenu comme décrivant le sens de la nouvelle occurrence du mot ambigu. En revanche, si **plusieurs associations** sont établies, celle ayant le score maximal est retenue.

Pour offrir au lecteur une vision générale du processus, on décrit tout d'abord le fonctionnement de l'algorithme de WSD (figure 3) puis, dans le paragraphe suivant, la fonction *renvoie_intersection_pondérée* est définie.

```
Paramètres : nouveau_contexte          // contexte de la nouvelle instance du mot ambigu M
             liste_clusters            // liste des clusters du mot ambigu M
Retourne : cluster_sélectionné         // cluster décrivant le sens de la nouvelle instance de M

traits_contexte = recupère_traits(nouveau_contexte)
liste_assoc_pondérées_clusters  Ø
cluster_sélectionné ← Ø

Pour chaque cluster de liste_clusters
      traits_cluster ← recupère_traits(cluster)
      liste_assoc_pondérées_clusters ← renvoie_intersection_pondérée(traits_contexte, traits_cluster)
FinPour
Si liste_assoc_pondérées_clusters non vide
     cluster_sélectionné ← renvoie_max(liste_assoc_pondérées_clusters)
     retourne cluster_sélectionné
Sinon
     retourne nil
FinSi
```

Figure 3. Algorithme de désambiguïsation lexicale

2.2.1.5. Pondération des associations

Toute association établie entre le nouveau contexte et un cluster d'EQVs est pondérée. Le **score** de l'association est calculé à l'aide des poids des traits qui la caractérisent relativement à chaque EQV du cluster. Rappelons que le poids (total) d'un trait *j* par rapport à un EQV *i* se calcule à l'aide du poids global de *j* et de son poids local par rapport à l'EQV *i* (cf . §2.3.4, chapitre 6).

Plus précisément : le poids de chaque trait relativement à chaque EQV du cluster est récupéré ; les poids sont ensuite additionnés et leur somme est divisée par le nombre de traits multiplié par le nombre d'EQVs, i.e. la moyenne des poids est calculée. L'association est pondérée par le score ainsi calculé.

La pondération de l'association entre le nouveau contexte et un cluster est effectuée par la fonction *trouve_intersection_pondérée*, décrite dans la figure 4.

```
Fonction trouve_intersection_pondérée

Paramètres :
    liste_traits_contexte          // liste des traits du contexte
    liste_traits_cluster           // ensemble des traits caractérisant un cluster
    EQVs_cluster                   // liste des EQVs d'un cluster
Retourne :
    association entre le couple <contexte, cluster> et le score attribué
    (association_contexte_cluster, score)

    compteur ← 0
    somme_poids ← 0
    score ← 0

    association_contexte_cluster ← traits_contexte ∩ traits_cluster
    Si association_contexte_cluster non vide
        Pour chaque trait de association_contexte_cluster
            Pour chaque EQV de EQVs_cluster
                poids_trait = renvoie_poids(trait, EQV)
                somme_poids += poids_trait
                compteur++
            FinPour
        FinPour
        score = somme_poids/compteur
        retourne <association_contexte_cluster, score>
    Sinon
        retourne nil
    FinSi
```

Figure 4. Pondération des associations entre contextes et clusters

Nous reprenons ici l'exemple fourni en paragraphe 2.2.1.3.

On the internal market there has been a standstill on many issues, from the free ***movement*** *of persons to the European company statute, to taxation, to the banking and insurance sector.*

Les associations établies par la comparaison entre le contexte de cette instance et les clusters de *movement* sont les suivantes :

a. {*διακίνηση, κινητικότητα, μετακίνηση*} : ***european***, ***person***
b. {*διακίνηση, κίνηση, κυκλοφορία*} : ***european***
c. {*διακίνηση, κίνηση, μετακίνηση* } : ***european***
d. {*κίνημα*} : ***european***

Les scores attribués à ces associations par la fonction '*trouve_intersection_pondérée*' sont décrits dans le Tableau 1. Dans le cas de l'association établie entre le nouveau contexte et le premier cluster, la somme des poids des 2 traits communs (*european*, *person*) par rapport à chacun des 3 EQVs (*διακίνηση, κινητικότητα, μετακίνηση*) est 6,539 ; cette somme est divisée par le produit du nombre de traits et du nombre d'EQVs (2 x 3 = 6). Ainsi, le score final attribué à l'association est 1,09. Les scores des autres associations sont calculés de la même manière.

1er cluster	**διακίνηση**	**κινητικότητα**	**μετακίνηση**	**somme**	**score**
european	1,955	0,696	1,393	6,539	1,09
person	1,103	0,696	0,696		
2e cluster	**διακίνηση**	**κίνηση**	**κυκλοφορία**	**somme**	**score**
european	1,955	1,393	3,951	7,299	2,433
3e cluster	**διακίνηση**	**μετακίνηση**	**κίνηση**	**somme**	**score**
european	1,955	1,393	1,393	4,741	1,58
4e cluster	**κίνημα**			**somme**	**score**
european	1,104			1,104	1,104

Tableau 1. Exemple de pondération des associations '*contexte-cluster*'

Dans le cas donc où de nombreuses associations sont établies pour le nouveau contexte, les clusters impliqués et les scores attribués aux associations sont sauvegardés dans un tableau. Le choix du sens véhiculé par la nouvelle instance du mot ambigu repose sur la sélection de l'association qui a le score le plus élevé. Le sens sélectionné pour l'exemple décrit ci-dessus est celui qui est décrit par le 2e cluster {*διακίνηση, κίνηση, κυκλοφορία*}, parce que l'association correspondante a obtenu le score maximal (2,433).

Il faut souligner que le **nombre** de traits communs entre le nouveau contexte et un cluster n'est pas aussi important que le score attribué à leur association. Ainsi, il se peut que le contexte partage un plus grand nombre de traits avec un cluster qu'avec un autre, mais que son association avec le premier soit plus faible.

Tel est le cas du 1er cluster (cf. Tableau 1) : son intersection avec le nouveau contexte contient plus de traits mais le score de cette association est inférieure à celui de l'association avec le cluster 2, qui partage moins de traits avec le nouveau contexte.

Nous donnons encore un exemple, où seulement deux associations sont créées entre le contexte de la nouvelle instance et les clusters de sens :

> *Article 56 of the Treaty, prohibiting currency speculation or any interference in the free movement of capital, must be removed so that damaging currency speculation may be checked by means of political control.*

L'ensemble des traits retenus de ce contexte est le suivant :

{*article*(NN), *treaty*(NN), *prohibit*(V), *currency*(NN), *speculation*(NN), *interference*(NN), *free*(JJ), *capital*(NN), *remove*(V), *damaging*(JJ), *currency*(NN), *speculation*(NN), *check*(V), *mean*(NN), *political*(JJ), *control*(NN)}.

Les associations créées mettent en contraste un des clusters décrivant le sens physique de *movement* (a) et le cluster décrivant son sens abstrait (b):

a. {*διακίνηση, κίνηση, κυκλοφορία*} : ***article, capital***

b. {*κίνημα*} : ***political***

1er cluster	**διακίνηση**	**κίνηση**	**κυκλοφορία**	**somme**	**score**
article	0,696	0,696	4,459	12,545	2,09
capital	1,399	2,324	2,971		
2e cluster	**κίνημα**			**somme**	**score**
political	0,748			0,748	0,748

Tableau 2. Exemple de pondération des associations '*contexte-cluster*'

Le sens attribué à cette instance est le sens décrit par le premier cluster, c'est-à-dire le sens de «movement physique».

2.2.2. Couverture de la méthode de WSD sur la base des clusters

Lorsque la méthode de WSD se base sur les traits caractérisant les clusters, un ensemble d'instances des mots ambigus ne se voit pas attribué de sens. Le nombre de traits caractérisant les clusters à plusieurs éléments est plus petit que celui de leurs traits assimilateurs. Cette différence est évidente dans la figure 1, où l'intersection des EQVs est plus restreinte que les ensembles décrivant leurs traits assimilateurs.

La faible quantité d'informations contenues dans l'intersection explique donc le fait que des associations ne soient pas toujours trouvées avec les contextes des nouvelles instances. Ainsi, la méthode de WSD utilisant les clusters propose un sens pour 73,65% des nouvelles instances des mots dont les traductions ont été manuellement repérées. Même si l'on ne tient pas compte des instances traduites dans le corpus de test par des EQVs rares[31], la couverture du système augmente très peu 74,79(%). La divergence observée au niveau des résultats est faible, en raison de la forte asymétrie des fréquences des EQVs. Cette asymétrie explique que l'élimination des EQVs rares d'un mot ambigu ne modifie pas de manière importante le nombre total de ses segments de test. En raison de l'impact très faible de l'élimination de ces EQVs sur les résultats, nous avons décidé de prendre en compte tous les EQVs des mots ambigus lors de l'évaluation et de ne pas filtrer les échantillons de test sur la base de la fréquence.

Nous estimons que le taux de couverture aurait été plus élevé si le corpus d'apprentissage avait été plus grand ou, si des cooccurrences d'ordre plus élevé avaient été utilisées. Davantage d'informations contextuelles auraient été alors disponibles, ce qui aurait permis le traitement d'un plus grand nombre de cas. Les solutions que nous avons envisagées pour augmenter le taux de couverture vont faire l'objet du paragraphe suivant.

[31] EQVs ayant une fréquence inférieure à 10 dans le corpus d'apprentissage.

2.2.3. Augmentation de la couverture de la méthode de WSD

2.2.3.1. Prise en compte des traits assimilateurs des EQVs clustérisés

Afin de compléter les résultats de la méthode de WSD et d'avoir des propositions pour des cas non traités sur la base des clusters, le nombre des ensembles de traits impliqués dans les comparaisons qui servent à désambiguïser les nouvelles instances peut être augmenté. Ainsi, hormis les traits caractérisant l'ensemble d'un cluster, les **traits assimilateurs** des paires des EQVs qui y sont inclus sont aussi exploités (dans le cas de clusters à plusieurs éléments). Etant donné que le nombre des traits assimilateurs des paires d'EQVs est souvent plus grand que celui des éléments de leur intersection (cf. §2.2.1.), l'exploitation de ces informations permet d'établir davantage d'associations avec les nouveaux contextes.

Dans le cas de repérage d'une association entre une nouvelle instance d'un mot et une paire d'EQVs, le cluster dans lequel la paire est incluse est considéré décrire le sens de l'instance en question. Ainsi, une partie des instances initialement non traitées sont désambiguïsées et le taux de couverture de la méthode augmente. Dans les cas où une paire se situe à l'intersection de clusters se chevauchant, les deux clusters sont proposés comme illustrant le sens de la nouvelle instance. La proposition des deux clusters correspond ici à la proposition d'un sens de granularité plus grossière, que les sens décrits par les clusters individuels. Si l'on souhaite que la granularité des sens proposés ne soit pas modifiée, il est également possible de sélectionner l'un des deux clusters, soit de manière aléatoire, soit en prenant en compte les poids des relations de la paire d'EQVs avec les autres membres des clusters en question. La sélection entre les deux clusters ne comporte pas le risque de sélectionner un sens distant, si l'on admet que lorsque l'intersection de deux clusters contient deux EQVs, une forte probabilité existe pour que les clusters décrivent des sens proches et qu'ils ne soient pas regroupés, faute d'informations contextuelles suffisantes (cf. §2.6.4, chapitre 6).

Dans le cas des mots dont les traductions sont manuellement repérées, la prise en compte des contextes assimilateurs des paires d'EQVs permet la

désambiguïsation de 15,86% des cas non traités par simple considération des clusters. Ainsi, le taux de couverture de la méthode atteint 89,51%. En revanche, dans le cas des mots du lexique bilingue automatiquement généré, la considération de ces contextes ne modifie que très faiblement le taux de couverture (augmentation de 1,86% ; couverture de 84,16%).

Cette différence observée au niveau de l'augmentation de la couverture entre lexiques manuellement et automatiquement générés s'explique par le fait que l'heuristique de l'utilisation des traits assimilateurs des EQVs s'applique seulement dans les cas de clusters composés de plus de deux éléments. Aucune modification de la couverture ne peut être observée dans les cas de « petits » clusters, à un ou deux éléments : dans le premier cas, les traits auxquels le nouveau contexte est comparé sont ceux qui caractérisent_l'unique EQV du cluster tandis que dans le deuxième cas, les traits assimilateurs des EQVs clustérisés sont exploités dès la première étape de la WSD. Par conséquent, l'application de cette technique sur les résultats obtenus à partir du lexique automatiquement généré ne provoque pas une forte augmentation de la couverture, parce que les clusters obtenus sur la base des informations de ce lexique sont plus petits et que le nombre de clusters de deux éléments est plus grand.

2.2.3.2. Prise en compte des traits spécifiques aux EQVs clustérisés

Une dernière heuristique prenant en compte les traits caractérisant séparément chacun des EQVs est appliquée dans le cas où il reste des instances non désambiguïsées, à la fin des deux premières étapes. Cette heuristique se fonde sur le fait que les ensembles de traits spécifiques à chaque EQV sont plus riches que ceux caractérisant les paires d'EQVs ou les clusters entiers[32].

Dans le cas de repérage d'une association entre une nouvelle instance du mot ambigu et un EQV, le cluster dans lequel l'EQV est inclus est considéré décrire le sens de l'instance en question. Comme dans le cas des paires d'EQVs, si l'EQV se situe à l'intersection de clusters se chevauchant, les deux clusters sont proposés comme illustrant le sens de la nouvelle instance, de granularité

[32] Les ensembles des traits spécifiques à chaque EQV sont ceux dont l'intersection constitue l'ensemble de traits assimilateurs d'une paire.

grossière. Néanmoins, la probabilité que les clusters ayant un seul EQV en commun décrivent des sens distants augmente par rapport au cas où l'intersection des clusters contient une paire d'EQVs. Ici aussi, il serait possible de sélectionner l'un des clusters qui se chevauchent, si une modification de la granularité des sens n'était pas souhaitée. Cette sélection est cependant risquée en raison de l'éventuelle ambiguïté de l'EQV partagé. En raison des inconvénients inhérents à cette heuristique, nous n'y avons eu recours qu'à la fin du processus de WSD, et seulement dans le cas où aucune solution n'avait été trouvée sur la base des clusters ou des paires d'EQVs. Cette heuristique augmente la couverture de 7,58%, dans le cas du lexique manuellement généré (atteignant ainsi 97,09%), et de 11,65%, dans le cas du lexique automatiquement généré (atteignant une couverture de 95,81%).

Les figures suivantes récapitulent les résultats de la couverture de la méthode en utilisant seulement les informations relatives aux clusters, en les combinant aux informations relatives aux paires d'EQVs, aussi bien qu'à celles qui caractérisent chacun des EQVs séparément.

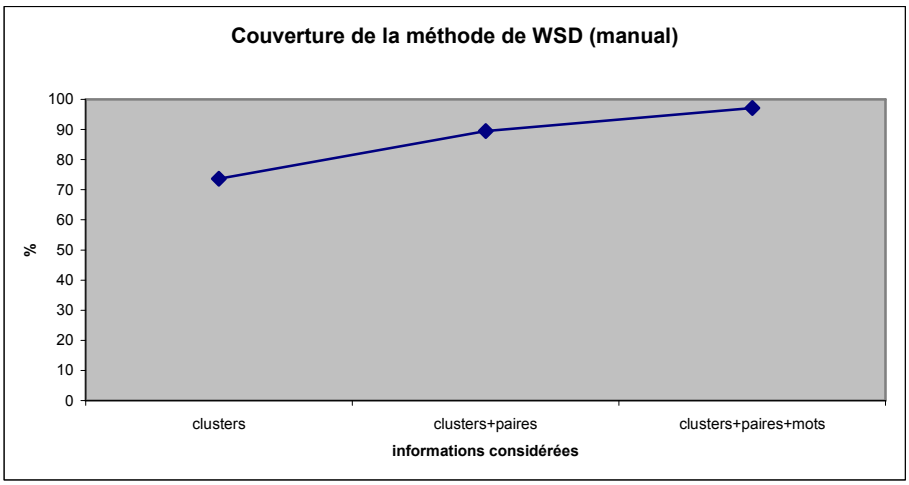


Figure 5. Couverture de la méthode de WSD (lexique manuellement généré)

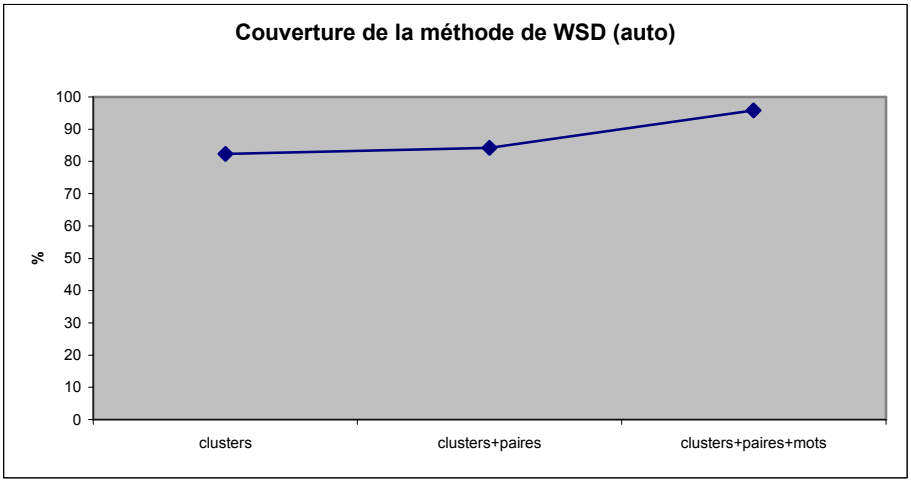


Figure 6. Couverture de la méthode de WSD (lexique automatiquement généré)

Au sein de la figure 5, l'apport de la considération des traits assimilateurs est évident ; il augmente la couverture de la méthode, couverture qui est augmentée davantage en prenant en compte les mots isolés. Au contraire, au sein de la figure 6, l'apport de la considération des traits assimilateurs est plus faible. La couverture de la méthode de WSD sur la base des informations liées aux clusters n'est pas considérablement modifiée lorsque les informations relatives aux paires sont prises en compte ; en revanche, une augmentation plus remarquable de la couverture est observée lors de l'exploitation des traits caractérisant les mots isolés.

Il est à noter ici que dans le cas où chaque EQV d'un mot ambigu est inclus dans un cluster distinct (c'est-à-dire qu'aucun regroupement n'a lieu), les prédictions de la méthode de WSD basées sur les informations relatives à ces EQVs sont comptées comme des prédictions au niveau des mots.

2.3. Exploitation des résultats de la désambiguïsation pour la sélection lexicale

La caractérisation du sens d'un mot source en contexte à l'aide d'un cluster correspond à la proposition d'un ensemble d'EQVs traduisant le sens. Le lien

établi de cette manière entre le mot et ses traductions possibles montre la relation existant entre les tâches de WSD et de sélection lexicale. Dans le cas de sélection d'un cluster à un EQV pendant la WSD, cet EQV est considéré comme traduisant l'instance en question ; l'étape de WSD coïncide ainsi avec celle de sélection lexicale.

En revanche, lorsque le sens sélectionné pour le mot source est décrit par un cluster à plusieurs éléments, le nombre des candidats de traduction du mot est restreint mais la sélection de la traduction n'est pas effectuée. Les EQVs clustérisés étant sémantiquement similaires, ils sont alors considérés comme plus ou moins substituables en tant que traductions de l'instance désambiguïsée du mot source.

Dans un cadre de Traduction Assistée par Ordinateur (Kay, 1980; Isabelle *et al.*, 1993 ; Piperidis *et al.*, 1999 ; Macklovitch *et al.*, 2000), les EQVs du cluster sélectionné lors de la WSD pour une instance d'un mot ambigu pourraient constituer des propositions de traduction sémantiquement pertinentes au niveau des mots. L'ensemble des candidats de traduction pour un mot serait limité, de cette manière, sur la base de critères sémantiques. En outre, au lieu d'être confronté à des solutions uniques, le traducteur aurait la possibilité de considérer des EQVs sémantiquement proches, véhiculant le bon sens du mot ambigu source. Le résultat de la WSD pourrait être ensuite raffiné par l'utilisateur du système de TAO en vue de sélectionner la traduction la plus appropriée pour l'instance du mot source en question.

Dans un cadre de TA, si la possibilité de révision des traductions produites par le système par un traducteur humain existait, la proposition d'alternatives pertinentes pourrait aider de manière importante le réviseur, chargé alors du choix final. En effet, un problème souvent cité par les traducteurs professionnels est que la révision des traductions « brutes » produites par les systèmes de TA est plus difficile et demande plus de temps que la traduction elle-même (Macklovitch, 1991). Une des raisons de ce constat est qu'il est difficile pour le traducteur de prendre de la distance par rapport à la proposition unique du système de TA, si cette suggestion est inappropriée, et de trouver une solution plus adéquate. La proposition faite au traducteur d'alternatives pertinentes pourrait donc faciliter son travail de révision.

Si cette révision du résultat de la TA n'est pas possible, le raffinement du résultat de la WSD, qui consiste à filtrer les EQVs du cluster proposé pour une instance du mot source, peut être effectué automatiquement par une méthode de sélection lexicale.

3. Sélection lexicale pour la traduction

3.1. Substituabilité des équivalents relativement aux contextes source et cible

Les EQVs clustérisés sont considérés, dans un premier temps, comme plus ou moins substituables en tant que traductions du mot source en contexte (cf. §3.3, chapitre 7). La possibilité de substitution des EQVs est estimée sur la base de leur similarité, induite par les contextes source du corpus d'apprentissage.

Il se peut pourtant que les EQVs d'un cluster ne soient pas tous adéquats et librement substituables dans la traduction et que leur utilisation ne soit pas toujours **naturelle**, même s'ils véhiculent le même **sens principal**. Les facteurs limitant la substituabilité des EQVs dans la traduction ont été analysés dans le paragraphe 4.2. du chapitre 7. La substituabilité effective des EQVs clustérisés en tant que traductions du mot source est estimée par référence au contexte de la LC dans lequel la traduction sera utilisée. C'est ce contexte qui « guide » la méthode de sélection lexicale dans le filtrage du cluster proposé par la WSD et le choix final de traduction. Bien que la proposition de la méthode de WSD se base sur les **relations de similarité** des EQVs par rapport aux contextes source, celle de la méthode de sélection lexicale dépend surtout de leurs **relations de dissimilarité** par rapport aux contextes cible.

Ainsi, tandis que les traits qui permettent de capter le noyau sémantique commun des EQVs (traits assimilateurs) sont suffisants pour l'étape de WSD, celle de sélection lexicale présuppose la prise en compte de leurs **différences**. Ces informations sont accessibles en prenant en considération les traits dissimilateurs de la LC. La possibilité d'accéder à des informations de type différent, selon les besoins qui se présentent au cours des deux étapes de traitement, informations permettant un raffinement plus ou moins grand des relations sémantiques

obtenues, permet de caractériser notre approche d'approche de **profondeur variable** (Kayser et Coulon, 1981)[33].

Nous faisons abstraction des facteurs extra-linguistiques pouvant influencer la sélection lexicale et des facteurs pertinents au niveau du document, en n'étudiant que l'apport du contexte lexical proche. Le contexte lexical de la LC sera appelé dorénavant **contexte de traduction**.

3.2. Contexte de traduction

3.2.1. Le contexte de traduction dans un cadre de TA

Le contexte de traduction est constitué des mots trouvés dans la traduction de la phrase d'entrée, la traduction du mot ambigu exceptée. Dans un système de TA, la constitution de ce contexte serait dépendante de l'approche de traduction adoptée (Specia *et al*., 2006a) : si, par exemple, le système traduit les mots dans l'ordre dans lequel ils apparaissent dans la phrase d'entrée, le contexte de traduction serait constitué des traductions des mots situés dans des positions antérieures au mot ambigu dans la phrase.

Alternativement, si le contexte (non-ambigu) du mot ambigu était traduit avant de procéder à sa désambiguïsation et à sa traduction, le contexte exploité pour la sélection de l'EQV le plus adéquat pour ce mot serait composé des traductions déjà fournies. D'autres variations seraient possibles pour des approches de TA traduisant soit des propositions, soit tous les mots simultanément, soit la phrase, en identifiant sa structure de base.

[33] En IA, les représentations de connaissances standard se situent à un niveau de profondeur fixe, c'est-à-dire que les objets manipulés sont décrits par une quantité d'informations qui reste constante pour chaque tâche. Néanmoins, le besoin d'exactitude de l'analyse computationnelle n'étant pas toujours le même pour chaque mot à traiter, on doit être capable de l'approfondir si besoin est. Ceci implique que chaque mot doit être lié à une quantité de connaissances ordonnée, qui se rendrait disponible de manière progressive lorsque la profondeur de l'analyse augmente. Il n'est alors pas nécessaire de rendre toutes ces informations disponibles dès le début. Kayser et Coulon proposent un traitement à « profondeur variable », qui utilise une description progressive des données, exploite des stratégies différentes en fonction de la qualité du résultat souhaité et contrôle continuellement cette qualité par une évaluation des approximations effectuées.

3.2.2. Exploitation du contexte de traduction par la méthode de sélection lexicale

Dans le cadre expérimental dans lequel notre travail se situe, le contexte de traduction est fourni par le **bitexte d'évaluation**, qui contient les traductions des phrases source. Le corpus parallèle permet en effet de simuler l'environnement qui serait fourni par un système de TA. L'utilisation de ce type de contexte dans des expériences sur l'évaluation de méthodes de WSD et de prédiction de traduction permet de prendre de la distance par rapport aux différentes approches et systèmes de TA et élimine le biais de la précision du système sur l'évaluation (Specia *et al*., 2006a).

Les unités de traduction constituant le **sous-corpus de test** d'un mot ambigu sont triées en fonction de la présence d'un de ses EQVs dans la phrase de la LC (cf. §2.2.5, chapitre 4), fournis par le lexique bilingue constitué lors de l'apprentissage. Un groupe d'unités est ainsi formé pour chaque EQV, tandis que les unités contenant plus d'un EQV sont éliminées[34]. L'EQV traduisant le mot source au sein d'une unité constitue la **traduction de référence** et, une fois repéré, il est remplacé par un **vide**. Soit l'unité de traduction correspondant au mot ambigu *movement* décrite dans la figure 7.

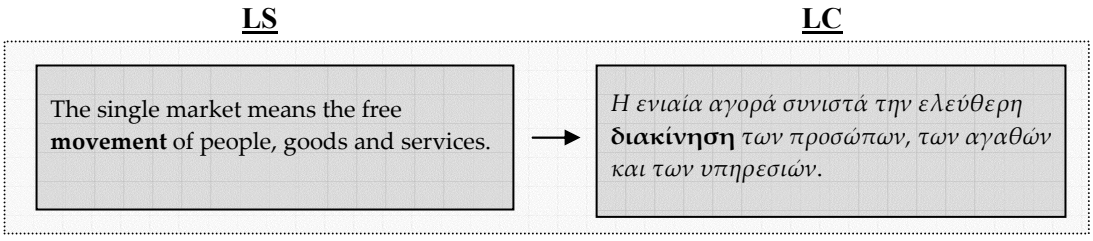


Figure 7. Unité de traduction du corpus de test

La traduction de référence, dans ce cas, est ***διακίνηση*** (diakinisi). Le vide qui remplace cet EQV indique la position que la traduction sélectionnée pour l'instance du mot source en question occupera dans le contexte cible, comme cela est montré dans la figure 8.

[34] Nous expliquons les raisons de cette élimination dans le paragraphe 1.3.11 du chapitre 4.

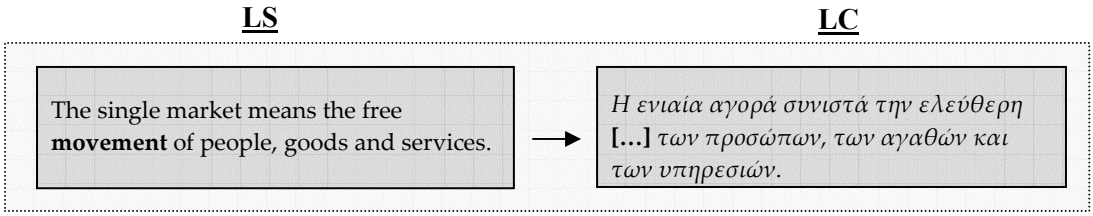


Figure 8. Unité de traduction comportant un "vide" dans le segment de la LC

Le rôle de la méthode de sélection lexicale est donc de remplir ce vide par la traduction la plus appropriée de l'instance du mot source en question, en exploitant le contexte de traduction. Le rôle de cette méthode pourrait être assimilé à celui du **modèle de langue** dans un système de TA, à la différence que ces modèles exploitent des informations locales (le plus souvent représentées à l'aide de *n*-grammes). En revanche, la méthode de sélection lexicale proposée profite d'un contexte plus large (l'ensemble du segment cible) qu'elle utilise dans une approche de **sac de mots**.

Cette conception du contexte de traduction ressemble à celle adoptée par Vickrey *et al*. (2005) pour résoudre ce qu'ils appellent un **problème simplifié de traduction**, qui consiste précisément en la prédiction de traduction pour un mot source polysémique en contexte. Specia *et al*. (2006a), dans les expériences concernant leur méthode de WSD pour la traduction, utilisent également le contexte de traduction constitué par les mots de la phrase cible, l'EQV du mot ambigu excepté. En fait, ils envisagent un système hypothétique de TA basé sur des règles, qui traduirait initialement tous les mots non-ambigus de la phrase source puis les mots ambigus (dans l'ordre dans lequel ils apparaissent dans la phrase) à l'aide du module de WSD. Leur contexte serait ainsi constitué des traductions de mots non-ambigus et éventuellement de celles de mots ambigus, qui auraient déjà été désambiguïsés.

3.2.3. Exploitation des informations de la LC

3.2.3.1. Traits de la LC retenus pendant l'apprentissage

L'identification des informations pertinentes de la LC relativement au filtrage automatique de la proposition de la WSD s'effectue par une analyse du contexte des EQVs dans les textes cible du corpus d'apprentissage. Ces informations consistent en les traits contextuels (noms, adjectifs et verbes) les plus fortement liés à chaque EQV.

L'apprentissage opère sur les **formes des mots** de la LC, à l'instar du processus d'apprentissage au niveau de la LS. La pertinence des traits est évaluée par leurs poids, calculés par rapport à chaque EQV lors du calcul de similarité au niveau de la LC (cf. §2.3.3, chapitre 6). Chaque élément retenu du contexte est pondéré en fonction de sa dispersion dans les textes du sous-corpus du mot ambigu (poids global) et de sa fréquence de cooccurrence avec chacun des EQVs (poids local). Le poids total d'un trait par rapport à un EQV – qui correspond au produit de son poids local par rapport à l'EQV et de son poids global – permet d'estimer sa pertinence.

Les informations des contextes cible retenues pendant l'apprentissage constituent une sorte de **référence** par rapport à laquelle les informations des contextes de traduction seront comparées.

3.2.3.2. Analyse du contexte de traduction

Les informations du contexte de traduction exploitées pour la sélection lexicale comprennent les formes des noms, adjectifs et verbes entourant le vide au sein de la phrase cible, prises en compte dans une approche de **sac de mots** (c'est-à-dire comme un ensemble de mots non ordonné). Pour que la comparaison des informations issues de nouveaux contextes aux informations retenues lors de l'apprentissage soit possible, le corpus de test a été lemmatisé et morphosyntaxiquement étiqueté[35]. Si nous reprenons le cas de l'unité de

[35] Le processus de lemmatisation et d'étiquetage morphosyntaxique de la partie grecque du corpus de test est décrit dans le paragraphe 2.2.3 du chapitre 4.

traduction de *movement* citée dans le paragraphe précédent, les éléments constituant le contexte de traduction sont les suivants :

{*ενιαίος* (AjBa), *αγορά* (NoCm), *συνιστώ* (VbMn), *ελεύθερος* (AjBa), *πρόσωπο* (NoCm), *αγαθό* (NoCm), *υπηρεσία* (NoCm)}[36]

Le contexte de traduction est donc constitué par les formes des mots qui y apparaissent. Cet ensemble de formes sera par la suite comparé aux traits retenus pour chacun des EQVs pendant l'apprentissage.

Le contexte de traduction peut être défini de différentes manières; par exemple, Specia *et al*. (2006a) combinent des *n*-grammes, constitués par des mots se trouvant à gauche et à droite de l'EQV dans la traduction (ce qu'ils appellent **contexte local**), et des sacs de mots de tous les mots du contexte de traduction (le **contexte du topic**). Dans cette étude, les *n*-grammes et les sacs de mots sont composés des traductions des mots déjà traduits de la phrase source et d'une des traductions possibles du mot ambigu[37].

Ayant exposé la manière dont le contexte de la LC est pris en compte lors de l'apprentissage et dans le cas de nouvelles occurrences des mots ambigus, dans le paragraphe suivant nous présenterons le fonctionnement de la méthode de sélection lexicale développée.

[36] Le jeu d'étiquettes utilisé pour le grec est décrit en détail en Annexe E2. Les champs des étiquettes qui nous intéressent ici sont ceux qui désignent la partie du discours, c'est-à-dire les deux premiers champs. Les verbes sont, par exemple, annotés sur 15 positions (« VbMnPpXxXxSgMaPePvNm ») ; les informations intéressantes de l'étiquette sont celles décrites par les deux premiers champs (« VbMn »), qui indiquent qu'il s'agit d'un verbe. De même, les noms sont annotés sur 5 positions (« NoCmFeSgAc »), et nous gardons seulement les informations décrites dans les deux premiers champs (« NoCm »).

[37] Les contextes constitués de cette manière sont ensuite utilisés pour effectuer des requêtes sur le Web (sur Google), ce qui permet d'explorer la fréquence de cooccurrence des mots formant ces contextes. L'EQV sélectionné est celui qui se trouve dans les contextes qui apparaissent le plus fréquemment, c'est-à-dire qui retournent le plus grand nombre de liens.

3.3. La méthode de sélection lexicale

3.3.1. Entrée de la méthode de sélection lexicale

La méthode de sélection lexicale s'applique **uniquement** dans le cas où la proposition de la méthode de WSD pour une instance d'un mot source concerne un cluster constitué de **plus d'un EQV**. Si un tel cluster est choisi comme illustration du sens d'une instance, il constitue alors l'entrée de la méthode de sélection, dont le rôle est de sélectionner parmi les EQVs du cluster celui qui serait le plus adéquat en tant que traduction de l'instance en question.

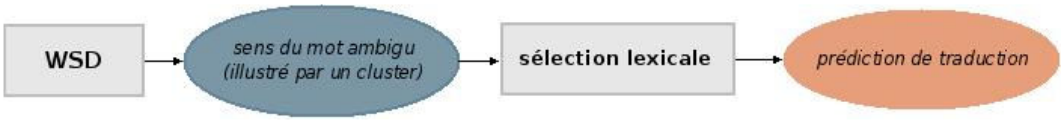


Figure 9. Entrée et sortie de la méthode de sélection lexicale

3.3.2. Adéquation des EQVs au sein du contexte cible

La méthode de sélection procède de la manière suivante :

1. elle analyse le contexte de traduction
2. elle compare les traits du nouveau contexte aux traits de la LC retenus, lors de l'apprentissage, pour chaque EQV du cluster proposé par la WSD
3. elle choisit l'EQV qui correspond à la traduction la plus adéquate de la nouvelle instance.

La comparaison des ensembles de traits permet de **pondérer** les EQVs relativement au contexte cible ; l'EQV sélectionné est ainsi celui qui obtient le score maximal. Cette pondération est conçue comme reflétant l'**adéquation** d'un EQV au sein du contexte de traduction. La sélection lexicale peut donc être

considérée comme étant guidée par des **préférences** d'emploi des EQVs clustérisés relativement au contexte cible.

Au contraire, dans le cas d'EQVs isolés formant un cluster à un seul élément et qui traduisent ainsi des sens distincts, la sélection de la traduction coïncide avec la sortie de la WSD. Dans ce cas, la sélection se base sur les contextes source et le recours aux informations de la LC n'est pas nécessaire. La traduction d'un sens par un seul EQV, dont l'emploi exclut celui des autres, permettrait de concevoir dans ce cas le processus de sélection comme obéissant à des **contraintes** plus strictes que dans le cas d'EQVs clustérisés.

Regardons un exemple où un cluster d'EQVs est sélectionné lors de la WSD pour une instance du mot *power*. La phrase d'entrée est la suivante:

*This means that the 'white gold', the water used in mountainous areas of Europe to generate hydroelectric **power**, will clearly acquire greater value as a fundamental resource for the mountain economy.*

Les clusters de *power* parmi lesquels la méthode de WSD doit choisir celui qui décrit le sens du mot dans ce contexte sont décrits dans le Tableau 3.

Clusters de *power*	**Description des sens**
δικαιοδοσία (dikaiodosia)	juridiction
αρμοδιότητα (armodiotita), *δύναμη* (dynami), *ισχύς* (ischys), *εξουσία* (eksousia)	pouvoir, autorité
ενέργεια (energeia), *ισχύς* (ischys), *δύναμη* (dynami)	puissance, force, énergie
ευχέρεια (efchereia)	pouvoir[38], droit
εξουσία (eksousia), *δυνατότητα* (dynatotita), *αρμοδιότητα* (armiodiotita)	pouvoir, possibilité
καθήκον (kathikon)	devoir

Tableau 3. Clusters de sens de *power*

D'après les résultats de la WSD, le sens de cette instance de *power* est décrit par le cluster {*ενέργεια*, *ισχύς*, *δύναμη*}. Ces EQVs présentent une similarité

[38] EQV utilisé très souvent pour traduire l'expression '*pouvoir d'appréciation*'.

sémantique, sans nécessairement être librement substituables au sein des contextes grecs. La méthode de sélection lexicale choisira donc parmi eux le plus adéquat pour traduire l'instance en question, sur la base du contexte de traduction. La traduction de la phrase d'entrée dans le corpus de test est la suivante :

> *Για τον «λευκό χρυσό», το νερό που χρησιμοποιείται στις ορεινές ζώνες της Ευρώπης για την παραγωγή υδροηλεκτρικής* ***ενέργειας****, το μέτρο αυτό σημαίνει μια προφανή ανάδειξη της αξίας του ως βασικού πόρου για την ορεινή οικονομία.*

d'où ce contexte de traduction est constitué :

> {*λευκός* (AjBa), *χρυσός* (NoCm), *νερό* (NoCm), *χρησιμοποιώ* (Vb), *ορεινός* (AjBa), *ζώνη* (NoCm), *Ευρώπη* (NoPr), *παραγωγή* (NoCm), *υδροηλεκτρικός* (AjBa), *μέτρο* (NoCm), *σημαίνω* (VbMn), *προφανής* (AjBa), *ανάδειξη* (NoCm), *αξία* (NoCm), *βασικός* (AjBa), *πόρος* (NoCm), *ορεινός* (AjBa), *οικονομία* (NoCm)}

La méthode de sélection lexicale sélectionne l'EQV *ενέργεια* pour cette instance de *power*, qui correspond à la traduction de référence.

CONCLUSION

Malgré la mise en question par Carpuat et Wu (2005a) du besoin de WSD pour la TA, les autres travaux qui explorent la question démontrent tous le rôle bénéfique de la WSD sur la qualité de la traduction. Dans ces travaux, la WSD est conçue comme l'exploitation de traits contextuels riches, non pris en compte par les systèmes de SMT. La distinction entre les sens des mots ambigus correspond à la distinction entre leurs EQVs, ce qui provoque l'assimilation, au sein des méthodes proposées, de la tâche de WSD avec celle de sélection lexicale. Cependant, comme nous l'avons déjà montré, une telle correspondance entre sens et EQVs peut être aisément remise en question sur un plan théorique.

La méthode de WSD que nous proposons exploite bien les EQVs de traduction des mots ambigus mais sans établir de correspondances biunivoques

entre sens et EQVs. En revanche, les sens proposés pour de nouvelles instances des mots ambigus correspondent aux clusters d'EQVs qui les décrivent. Ces clusters sont constitués d'un ou plusieurs éléments. Dans le premier cas, la proposition de l'WSD pour une instance peut être directement considérée comme la traduction de l'instance en question. Dans le cas de clusters à plusieurs éléments, la sélection de l'EQV de traduction le plus adéquat en contexte se base sur une méthode de sélection lexicale, qui exploite le contexte de la traduction.

Cette manière de concevoir les relations entre sens et EQVs permet un traitement différencié des EQVs en fonction de leur statut : les EQVs sémantiquement similaires (clustérisés) sont considérés comme plus ou moins substituables en tant que traductions d'un sens du mot source ; au contraire les EQVs formant des clusters à un seul élément sont considérés traduire des sens clairement distincts. La sélection entre les EQVs clustérisés se caractérise ainsi par une plus grande **flexibilité** et s'effectue sur la base de **préférences** concernant leur utilisation dans le texte cible ; en revanche, la sélection d'un EQV isolé peut être conçue comme imposée par une **contrainte sémantique**. La prise en compte des relations de similarité entre les EQVs présente des avantages tel que la possibilité de propositions multiples sémantiquement pertinentes au niveau des mots dans un cadre de TAO. Elle rend aussi possible la pénalisation différenciée des erreurs de traduction par les métriques d'évaluation, sans recours à des ressources extérieures. L'exploitation de ces relations pourrait même s'avérer profitable dans un cadre de recherche d'information multilingue, dans la mesure où une plus grande quantité d'informations sémantiquement pertinentes pourraient être proposées à l'utilisateur.

L'évaluation quantitative des méthodes de WSD et de sélection lexicale proposées sera présentée dans le chapitre 9. Mais avant cela, nous allons exposer les manières dont nous avons procédé à l'évaluation qualitative des résultats de notre méthode d'acquisition de sens.

CHAPITRE 9

EVALUATION QUALITATIVE DES SENS ACQUIS AUTOMATIQUEMENT

INTRODUCTION

Dans les chapitres précédents, nous avons analysé les principes de fonctionnement de notre méthode d'acquisition de sens et avons présenté une partie des résultats obtenus. Puis, nous avons montré comment ces résultats peuvent être exploités pour la WSD et la sélection lexicale dans des applications relatives à la traduction. Après avoir ainsi montré le lien unissant ces méthodes, nous allons désormais nous attacher à présenter les étapes d'évaluation qualitative effectuées sur les résultats de l'acquisition de sens. Nous avons déjà souligné la difficulté de validation de sens induits à partir de données textuelles, faute d'un étalon d'or en sémantique lexicale et de critères objectifs d'estimation de la validité des sens proposés (cf. §3, chapitre 2).

Notre évaluation qualitative consiste à comparer nos résultats aux descriptions fournies par une autre méthode d'analyse sémantique basée sur les données. Il s'agit de la méthode des **Miroirs Sémantiques** (Dyvik, 1998a, 2003, 2005), méthode qui exploite uniquement des informations traductionnelles (et non des informations contextuelles). Le fonctionnement de cette méthode sera d'abord décrit, puis un échantillon des résultats obtenus par son application à nos données sera présenté. La similarité des résultats obtenus par ces deux méthodes, fondées sur des informations de nature différente, nous servira alors d'indice de validité de nos résultats.

L'autre étape de cette évaluation qualitative consistera à comparer les descriptions sémantiques fournies par notre méthode à celles trouvées dans une ressource lexico-sémantique existante, le réseau multilingue de wordnets BalkaNet (Tufiş *et al*., 2004a).

1. Comparaison des résultats de deux méthodes d'analyse sémantique

1.1. La méthode des Miroirs Sémantiques

1.1.1. Analyse sémantique des mots à l'aide de leurs *reflets*

La méthode des **Miroirs Sémantiques** (Dyvik, 1998a ; 2003 ; 2005) est l'une des premières **méthodes traductionnelles** d'analyse sémantique élaborée. Cette méthode s'appuie sur des informations provenant d'un corpus parallèle aligné au niveau des mots et ne nécessite pas de ressources lexicales ou sémantiques extérieures. Chaque langue représentée dans le corpus est traitée comme le « miroir sémantique » de l'autre. L'analyse des **reflets** des mots dans leur miroir, correspondant à leurs traductions, rend possible le repérage de leurs sens et l'analyse des relations sémantiques qui existent éventuellement entre eux. La méthode des Miroirs exploite donc uniquement des **relations traductionnelles** ; son entrée est constituée par les traductions alternatives fournies pour chaque forme de mot dans les résultats de l'alignement lexical d'un corpus parallèle.

Les traductions alternatives d'un mot, étant liées à ses sous-sens (ou à ses aspects sémantiques) différents, sont considérées comme révélant une manière de diviser son potentiel sémantique. Les sous-sens du nom allemand *Hexe*, par exemple, qui peut être traduit en anglais par *hag* et *witch*, correspondent, de cette manière, aux paires de mots ordonnées <Hexe, hag> et <Hexe, witch>[1]. Ces paires d'éléments peuvent être assimilées à des **traits sémantiques** attribuables aux unités lexicales dont elles sont dérivées, ainsi qu'à celles pouvant en **hériter**. Ces traits encodent donc des sous-sens **partagés** par les unités lexicales et peuvent, ainsi, servir de **mécanismes classificatoires** à leur regroupement.

La sortie de la méthode consiste en un **réseau lexico-sémantique** complexe similaire à WordNet. Au sein de ce réseau, les relations entre les sens lexicaux au niveau monolingue sont modélisées et les sens sont mis en correspondance avec ceux repérés dans la langue miroir. Ces correspondances relient les vocabulaires des deux langues.

1.1.2. Hypothèses théoriques sous-jacentes à la méthode des Miroirs

Le réseau ainsi construit permet de traiter chaque langue comme le miroir sémantique de l'autre, sur la base des hypothèses suivantes (Dyvik, 2003, 2005) :

i. les ensembles de traductions de mots sémantiquement proches présentent généralement un fort degré de recouvrement[2]
ii. les mots ayant des sens larges ont tendance à avoir un plus grand nombre de traductions que les mots ayant des sens plus restreints
iii. si le mot *a* est un hyponyme du mot *b*, les traductions possibles de *a* constitueront probablement un sous-ensemble des traductions possibles de *b*
iv. l'ambiguïté contrastive caractérisant les mots dont deux sens ne sont pas liés est une propriété historiquement accidentelle et

[1] Ces ensembles seraient plus importants si plusieurs langues étaient prises en compte simultanément.
[2] Pour Resnik (2004) une traduction commune (par ex. *rive*) dans la LC est suffisante pour considérer que les mots de la LS (par ex. *bank* et *shore*) partagent un élément de sens.

idiosyncrasique, dont il ne faut pas s'attendre à trouver des instances dans d'autres langues[3]

v. des mots ayant des sens non liés n'ont pas les mêmes traductions dans une autre langue, sauf lorsque la traduction commune est ambiguë de manière contrastive entre les deux sens.

Ayant présenté les principes de la méthode, nous allons désormais présenter son application sur les données de notre corpus d'apprentissage.

1.2. Analyse de nos données traductionnelles à l'aide des Miroirs

1.2.1. Le but de l'analyse

L'intérêt d'appliquer la méthode des Miroirs à nos données est triple :

a. évaluer les distinctions sémantiques obtenues par notre méthode d'acquisition de sens – basée sur des informations à la fois contextuelles et traductionnelles – par comparaison avec les distinctions fournies par une méthode basée sur des informations de type différent (qui sont, elles, uniquement traductionnelles)

b. analyser la sémantique des EQVs clustérisés, ainsi que leurs relations, afin d'explorer les correspondances entre les régions sémantiques des mots des deux langues.

Notre méthode d'acquisition de sens éclaire la sémantique des mots ambigus en révélant des distinctions sémantiques sur la base de motifs contextuels statistiques, et propose des correspondances inter-langues de granularité variée entre les sens de ces mots et les clusters de leurs EQVs. Ces correspondances étant souvent de type « un à plusieurs », la méthode des Miroirs servira à étudier les relations entre les différents EQVs qui décrivent un sens. En ce qui concerne le point (iii) cité ci-dessus, notre méthode directionnelle permettant l'analyse de la sémantique des EQVs qui est liée à celle des mots

[3] Sauf dans le cas où une évolution parallèle s'est produite dans des langues apparentées.

source, les Miroirs nous donneront la possibilité d'avoir une image plus claire de la polysémie caractérisant les EQVs.

1.2.2. Données traductionnelles utilisées

1.2.2.1. Lexiques automatiquement construits

La méthode des Miroirs a été initialement appliquée sur les données de nos lexiques automatiquement générés ; le processus de construction de ces lexiques est décrit en paragraphe 1.3.3. du chapitre 4. Plus précisément, les données utilisées sont celles issues des lexiques bilingues dans les deux directions qui concernent les **noms**, obtenus après filtrage par partie du discours (cf. §1.3.4, chapitre 4) et par intersection des associations de traduction (cf. §1.3.5, chapitre 4).

1.2.2.2. Neutralisation de l'effet de la direction de traduction

L'analyse des résultats de l'alignement automatique a rendu évidentes des « lacunes » au niveau des correspondances extraites du bitexte qui, toutefois, paraissent justifiées lors du travail sur des textes réels. Il peut, en effet, arriver qu'une correspondance de traduction soit repérée entre un mot de la LS et un mot de la LC (par exemple, 'a→b'), sans que ces mots soient mis en correspondance lors de l'inversion de la direction de traduction (c'est-à-dire sans que la correspondance 'b→a' soit repérée). Pourtant, le non repérage de cette relation n'exclut pas la possibilité d'une correspondance de traduction entre les deux mots, voire même l'existence d'une telle correspondance dans le corpus, qui demeure « cachée » en raison d'autres facteurs.

Ce type de « manques » constitue un problème pour l'application de la méthode des Miroirs sur les données de l'alignement. Un des moyens d'y remédier consiste à **généraliser** les correspondances de traduction repérées dans le corpus, indépendamment de la direction de traduction. Ainsi lorsqu'un mot *a* fait partie de l'ensemble d'EQVs de *b*, *b* est considéré comme un EQV possible de *a*, même si leur relation n'est pas repérée dans le corpus. Ainsi, l'absence de *b* de

la liste d'EQVs de *a* est considérée comme étant due à des raisons autres que traductionnelles ou sémantiques.

Cette hypothèse peut néanmoins être remise en cause sur un **plan théorique**. La spécificité des systèmes linguistiques ne permet pas de supposer l'application universelle de la **bi-directionnalité** des relations de traduction. Ainsi, même si un mot de la LC est un bon EQV pour un mot de la LS dans une direction, il se peut qu'un autre mot soit plus adéquat à la traduction de l'EQV lors de l'inversion de la direction de traduction. Cette question touche donc à la question de l'adéquation des EQVs en tant que traductions des mots, propriété qui peut être conçue comme graduelle.

Sur un **plan pratique**, l'inversion de la direction de traduction peut donc impliquer la définition de **préférences de traduction** différentes. Ce qui paraît relativement évident dans une direction peut exiger, dans l'autre direction, la prise en considération de contraintes supplémentaires dans un système de TA, liées à la spécificité des systèmes linguistiques (Isabelle, 1989). Si, par exemple, une distinction sémantique est effectuée par deux mots dans une langue et non dans une autre, les mots seront alors traduits dans la deuxième langue par un mot plus général englobant leurs sens. Lorsque ce mot générique doit, en retour, être traduit dans la première langue, une sélection doit être effectuée entre les traductions spécifiques. Le besoin de définir des préférences de traduction s'impose donc pour l'une des deux directions[4].

Malgré ce besoin de prise en compte des spécificités, la construction de systèmes réversibles de traduction demeure possible. Il suffit que les spécificités en question soient bien décrites au niveau de la grammaire et du dictionnaire utilisés pour chaque langue. Pour qu'un système de TA soit réversible, il faut donc que toutes ses composantes le soient et que l'analyse et la génération pour

[4] Par exemple, les pronoms personnels français *tu* et *vous* sont tous les deux traduits en anglais par *you* mais, lors de l'inversion de la direction de traduction, il faudra plus d'indices pour effectuer un choix. De telles contraintes sont souvent également nécessaires au niveau de la grammaire. Par exemple, dans la traduction de l'anglais vers le français, le présent simple et le présent progressif (continu) peuvent tous les deux être traduits par le présent simple en français. Dans ce sens, le problème paraît simple et une analyse plus profonde est inutile. Cependant, dans une perspective de traduction du français vers l'anglais, un certain nombre d'indicateurs contextuels doivent être pris en compte, qui définissent si le présent simple en français doit être traduit en anglais par le présent simple ou le présent progressif (Isabelle, *ibid*.).

une langue donnée se fondent sur la même grammaire et le même dictionnaire (Dymetman et Isabelle, 1988).

Dans notre travail, la question de la réversibilité se pose au niveau du lexique des deux langues, c'est-à-dire au niveau des correspondances lexicales dans un corpus parallèle. Nous pourrions dire qu'il s'agit, là, plutôt d'une manière **statique** de concevoir les relations de traduction entre les unités lexicales, et non **dynamique**, qui impliquerait une sélection, comme c'est le cas dans les systèmes de TA. La généralisation effectuée au niveau des résultats de l'alignement des mots peut aussi être vue comme une **neutralisation** de l'effet de la direction de traduction. Cette neutralisation est en accord avec une des hypothèses de base de la méthode des Miroirs Sémantiques, qui défend la **symétrie** de la relation de traduction (Dyvik, 2003, 2005).

En acceptant donc cette hypothèse dans ce cadre bien précis, nous supposons que le manque d'informations est davantage lié à des facteurs qui influencent le résultat du calcul statistique qu'à des raisons sémantiques. Un de ces facteurs peut être la différence de fréquence et de dispersion des mots dans le corpus, qui influe sur le calcul de la probabilité de leur association. Sur un plan pratique, la décision d'effectuer cette généralisation implique que chaque fois qu'un mot d'une L2 est inclus dans l'ensemble d'EQVs d'un mot de la L1, ce mot est automatiquement inclus parmi les EQVs du mot de la L2, même s'il n'y figure pas à l'issue du processus d'alignement des mots. En adoptant cette généralisation, nous décidons donc d'ignorer les informations de direction de traduction[5].

Ce processus de généralisation a été initialement appliqué aux résultats de l'alignement lexical, filtrés en fonction des parties du discours (étape 4), et avant leur filtrage par intersection. Cependant, cette généralisation a constitué une source supplémentaire d'erreurs, augmentant le **bruit** au sein des résultats. Le filtrage en fonction des parties du discours n'étant parvenu, en effet, à éliminer qu'une petite partie du bruit des résultats de l'alignement des mots (non validés

[5] Nous parlons ici de L1 et de L2, et non de LS et LC, en raison de la neutralisation de l'effet de la direction de traduction. Ce type de notation sert seulement à illustrer les inversions de la direction de traduction. Les caractérisations source et cible n'ont pas d'importance lorsque les informations de direction de traduction sont ignorées.

à la main)[6], dans les cas de correspondances erronées dans l'une des deux directions, ces correspondances se sont alors retrouvées dans l'autre direction à l'issue de la généralisation. Ainsi ce processus a contribué à la **propagation d'erreurs** dans les résultats.

Mais, le filtrage des lexiques des deux directions sur la base de leur intersection, qui vise à préserver les associations issues des deux alignements (direct et inverse), élimine, quant à lui, une grande partie du bruit. Ne retenir, effectivement, que l'intersection de deux ensembles d'associations implique que les mêmes correspondances soient trouvées dans les deux directions. Le filtrage par intersection permet donc et de filtrer une grande partie du bruit et de neutraliser la direction de traduction.

Les données qui résultent de ces filtrages ont été traitées mais ne se sont pas avérées finalement suffisantes pour une description complète, par la méthode des Miroirs, de la sémantique des mots inclus dans ce lexique. La raison en est le **bas rappel** caractérisant les données ayant subi des filtrages consécutifs, dans le but d'en éliminer le bruit introduit par l'alignement. Ce faible taux de rappel concerne le nombre de traductions possibles fournies pour les mots dans ces lexiques. Les résultats du traitement de ces données n'ont donc pas été exploités finalement, puisque leur qualité ne permettait pas une comparaison intéressante avec les résultats obtenus par notre méthode d'acquisition de sens.

1.2.2.3. Lexiques bilingues manuellement construits

Contrairement aux données des lexiques automatiquement générés, les données du lexique anglais-grec que nous avons créé à la main sont caractérisées par une précision et un rappel les plus élevés possibles. Ces données ont été utilisées lors du développement de notre méthode d'acquisition de sens. Pour en comparer les résultats à ceux des Miroirs, nous avons choisi d'en limiter l'application à cet ensemble de données.

[6] En raison aussi de quelques erreurs au niveau de l'étiquetage morpho-syntaxique du corpus d'apprentissage qui n'a pas été validé à la main.

Les Miroirs nécessitant des données d'entrée dans les deux directions de traduction, nous avons construit manuellement un lexique dans la direction grec-anglais. Plus concrètement, nous avons inversé la direction de traduction **trois fois**, comme cela est décrit dans le paragraphe 1.3.9. du chapitre 4 (anglais-grec, grec-anglais, anglais-grec, grec-anglais). Le lexique anglais-grec généré suite à ces inversions contient 248 entrées, et le lexique grec-anglais 661. Deux lexiques supplémentaires ont été créés, ne contenant pas les EQVs utilisés une seule fois pour traduire les mots source, c'est-à-dire les *hapax*. Le temps requis pour l'élaboration de ces lexiques explique que nous ayons restreint l'application des Miroirs à un sous-ensemble des mots analysés par notre méthode d'acquisition de sens.

En utilisant ces informations de traduction, la méthode des Miroirs établit :

- des **listes complètes de thesaurus** contenant des informations sémantiques pour tous les mots dans les fichiers d'entrée
- des **entrées individuelles de thésaurus** et des **semi-treillis sémantiques**, décrivant les sens et les sous-sens des mots ainsi que leurs relations de synonymie, d'hyponymie et d'hyperonymie. Les semi-treillis illustrent graphiquement le réseau de relations sémantiques entre les sens.

1.3. Fonctionnement de la méthode des Miroirs

1.3.1. Individuation des sens

La première étape de l'application de la méthode des Miroirs consiste à distinguer les sens des mots ambigus[7]. L'ensemble des traductions dans une langue L2 d'un mot *m* d'une L1 est appelé le **premier *t*-image** de *m*. Le premier *t*-image du mot *plant*, par exemple, est constitué de l'ensemble de ses traductions

[7] Sur la base des hypothèses (iv) et (v) (voir paragraphe 1.1.2 de ce chapitre).

en grec : {*φυτό*, *εγκατάσταση*, *σταθμός*, *εργοστάσιο*, *μονάδα*, *εργαστήριο*, *φυτάριο*}[8].

En inversant la direction de traduction (L2→L1), les traductions des membres du premier *t*-image de *m* forment des ensembles de mots dans la L1. La totalité de ces ensembles est appelé le ***t*-image inverse** de *m*. Dans le cas de *plant*, le *t*-image inverse est constitué des ensembles de traductions en anglais de chacun de ses EQVs grecs, comme illustré dans la figure 1[9].

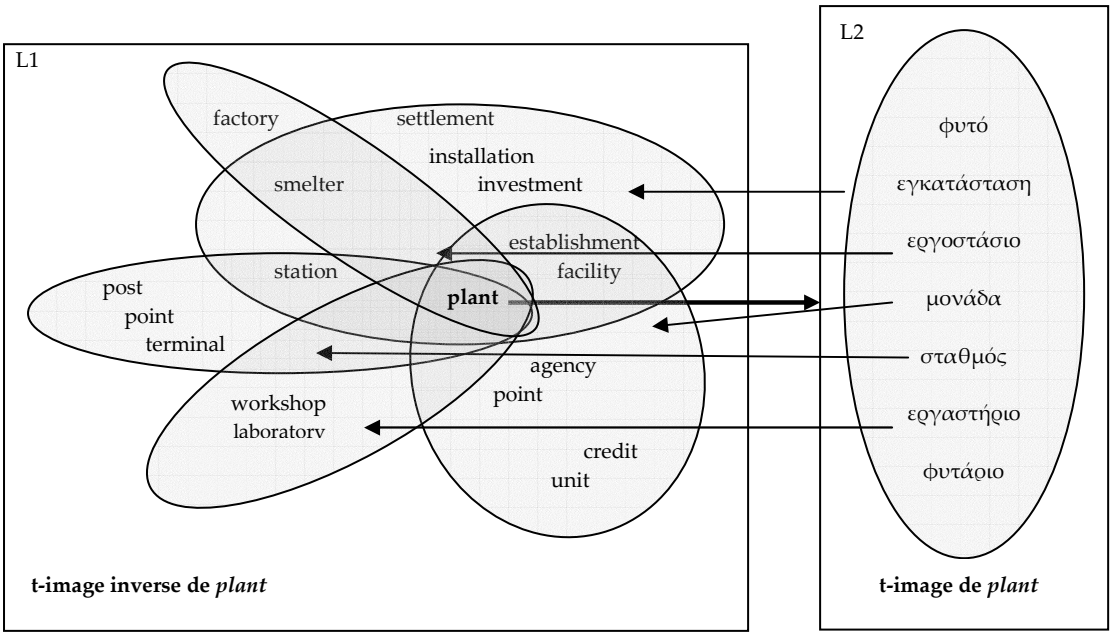


Figure 1. Le premier *t*-image et le *t*-image inverse de *plant*

Les ensembles de mots du *t*-image inverse (dans la L1) présentent des **recouvrements**. Lorsque l'intersection de deux ensembles contient plus d'un élément, ces ensembles peuvent être **regroupés**. En revanche, l'existence d'un seul mot à l'intersection de deux ensembles ne constitue pas un indice suffisant pour le regroupement car il se peut que le mot commun soit ambigu entre des sens distincts représentés par les deux ensembles. Par exemple, le mot ambigu (*plant*) se trouve à l'intersection de tous les ensembles de traductions des mots de

[8] L'EQV « *φυτάριο* », très rare, n'est pas pris en compte par notre méthode d'acquisition de sens parce qu'il se trouve, dans le corpus d'apprentissage, au sein d'unités de traduction contenant plus d'un EQV du mot *plant*, qui ont été éliminées.

[9] Pour rendre le schéma plus lisible, nous n'y avons pas inclus les relations de *φυτό* et de *φυτάριο*, dont le seul EQV en anglais est le mot *plant*.

son premier *t*-image, mais cela ne constitue pas un indice suffisant pour leur regroupement : l'intersection doit contenir des mots supplémentaires afin de pouvoir en tirer des conclusions.

Chaque groupe de mots formé dans la L1, en tenant compte des informations sur les intersections, correspond à un **sens distinct** du mot ambigu. Les mots de la L2 correspondant à ces groupes sont considérés comme **sémantiquement similaires**. Par exemple, l'image de *εγκατάσταση* (egkatastasi) dans la L1 recoupe celle de *σταθμός* (stathmos) (intersection : *plant*, *station*) ; celle de *μονάδα* (monada) (intersection : *plant*, *establishment*, *facility*) et celle de *εργοστάσιο* (ergostasio) (intersection : *plant*, *smelter*). Ces quatre ensembles de mots sont donc regroupés dans la L1 et les mots de la L2, *εγκατάσταση*, *σταθμός*, *μονάδα* et *εργοστάσιο*, sont alors considérés comme sémantiquement similaires et sont également regroupés. Les groupes créés dans le premier *t*-image du mot ambigu, par **projection** des groupes formés dans le *t*-image inverse, décrivent ses **partitions sémantiques**. Ainsi, les sens du mot ambigu sont identifiés et associés chacun à son propre premier *t*-image. Les partitions du premier *t*-image de *plant* (figure 2) décrivent les quatre sens du mot.

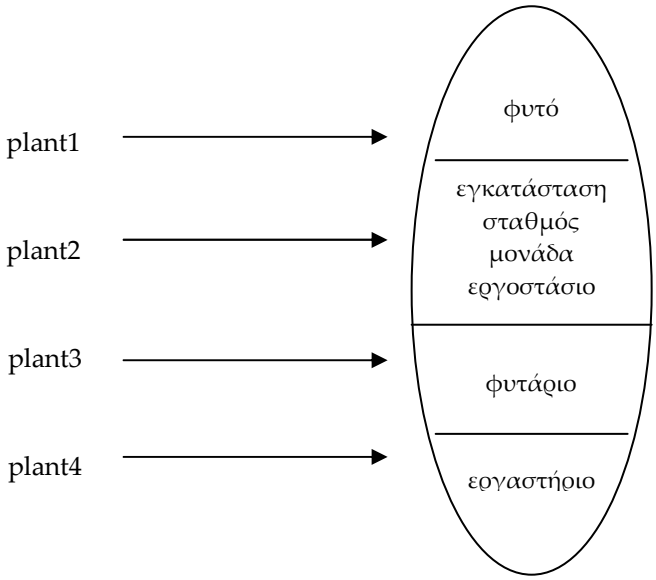


Figure 2. Partitions sémantiques de *plant*

Une inversion de la direction de traduction (L2→L1) suffit donc pour identifier les partitions sémantiques au sein du premier *t*-image d'un mot ambigu de la L1[10].

Afin d'analyser la sémantique des EQVs des mots ambigus, nous avons procédé à deux autres inversions de la direction de traduction (cf. §1.3.9, chapitre 4). La deuxième inversion (L1→L2) donne un ensemble de *t*-images dans la L2 de tous les membres de l'**union du *t*-image inverse** du mot ambigu, appelé son **deuxième *t*-image**[11]. La troisième inversion de la direction de traduction consiste à repérer des correspondants anglais de tous les mots de la L2 repérés lors de l'étape précédente et dont les correspondants n'avaient pas été repérés lors de la première inversion[12]. Les trois inversions de la direction de traduction permettent ainsi d'analyser les relations de traduction d'un grand nombre de mots des deux langues.

1.3.2. Création de champs sémantiques

Les sens des mots des deux langues, repérés de la manière décrite ci-dessus, peuvent être regroupés en **champs sémantiques**. Un champ sémantique est un **ensemble de sens**, directement ou indirectement liés entre eux par une **relation de proximité sémantique**. Dans une approche traductionnelle, les champs sémantiques sont définis sur la base des *t*-images qui se chevauchent : deux sens appartiennent au même champ sémantique,

- si au moins un sens dans l'autre langue correspond aux deux sens d'un point de vue traductionnel, c'est-à-dire si leurs premiers *t*-images se recouvrent[13] ou,
- s'il existe une séquence de telles *t*-images les liant entre eux (Dyvik, 2003 ; 2005).

[10] L'identification de ces partitions peut également se faire au sein du *t*-image inverse.

[11] Pendant cette étape, les équivalents de tous les mots anglais repérés au cours de l'étape précédente sont repérés dans le corpus (par ex. *smelter* : {*εργοστάσιο, εγκατάσταση*}, *installation* : {*συσκευή, εγκατάσταση*} etc.)

[12] Par exemple : *συσκευή* {*apparatus, pen, appliance; equipment, installation, device*}, etc.

[13] Après l'identification des sens, un seul élément situé à l'intersection des premiers *t*-images suffit.

La correspondance de traduction étant traitée comme une relation symétrique (la direction de traduction n'est pas prise en compte), des champs sémantiques **associés** sont obtenus dans les deux langues. Chaque champ d'une paire (*c1* et *c2*) impose une **structure de sous-ensembles** à l'autre, étant donné que tous les *t*-images des éléments de *c1* sont des sous-ensembles de *c2* et *vice versa*. Cette structure de sous-ensembles permet de dériver des informations riches sur les relations entre les sens, d'après les hypothèses (i)-(iii) des Miroirs (cf. §1.1.2.).

Ainsi, si un sens appartient à plusieurs *t*-images, cela signifie qu'il a de nombreux « partenaires » de traduction au sein du champ associé de l'autre langue. Ces sens sont supposés être plus **larges**, selon l'hypothèse (ii) des Miroirs[14], que d'autres sens du champ. En outre, si deux sens co-apparaissent dans plusieurs sous-ensembles, cela signifie qu'ils partagent plusieurs traductions et qu'ils doivent donc être **sémantiquement liés**. Les structures de sous-ensembles contiennent des informations riches sur les relations sémantiques inter-sens, qui peuvent être encodées dans des **représentations sémantiques** sous la forme d'**ensembles de traits** associés aux sens.

1.3.3. Construction de représentations sémantiques

Des représentations sémantiques sont d'abord construites pour les **sommets**, paires de sens qui sont à la fois :

- liés du point de vue de la traduction
- membres du plus grand nombre de sous-ensembles.

Dans le cas de *plant*, le sommet est constitué par la paire de sens *plant1* et *εγκατάσταση1*. Un trait est élaboré à partir de ces sens ([plant1|εγκατάσταση1]) qui est, ensuite, attribué aux deux sens et hérité par tous les sens situés au-dessous dans la représentation. Dans notre exemple, il s'agit des sens situés sous *plant1* dans le premier *t*-image de *εγκατάσταση1* et de ceux situés sous

[14] Hypothèse ii : « Les mots avec des sens larges ont tendance à avoir un plus grand nombre de traductions que les mots avec des sens plus restreints.».

εγκατάσταση1 dans le premier *t*-image de *plant1*. Il s'agit là de paires de sens liés du point de vue de la traduction mais membres d'un nombre de sous-ensembles plus petit que celui des sous-ensembles auxquels appartiennent les sommets. Puis, la procédure se répète pour les plus hauts sommets suivants. Un sens ne transfère que ses propres traits aux sens situés au-dessous de lui et non les traits dont il a lui-même hérités. A la fin de ce processus, des ensembles de traits sont assignés à tous les sens des champs des deux langues.

Dorénavant, les **relations hiérarchiques** au sein d'un champ sont exprimées à l'aide de relations d'**inclusion** et de **chevauchement** entre leurs ensembles de traits : les sens larges sont caractérisés par de petits ensembles de traits, tandis que leurs hyponymes sont représentés par des surensembles de ces ensembles de traits[15]. Les ensembles de traits forment ainsi un **semi-treillis** (supérieur), qui peut être décrit à l'aide d'un **graphe**. Les sens représentés par les nœuds « dominateurs » (supérieurs) sont des hyperonymes des sens des nœuds « dominés » (inférieurs).

Un semi-treillis supérieur est un ensemble partiellement ordonné, dans lequel chaque paire d'éléments a une borne supérieure (*least upper bound*). Pour chaque paire d'ensembles de traits, soit l'un inclut l'autre, soit il existe un élément qui consiste en l'intersection des deux ensembles. Ainsi, si deux ensembles de traits se recouvrent sans inclusion, un **nœud mère virtuel** est construit (*x-nœud*), représentant leur intersection. Dans la figure 3, nous décrivons le semi-treillis du champ sémantique du sens industriel de *plant*. Contrairement aux nœuds virtuels, les **nœuds non-virtuels** sont étiquetés par des formes lexicales qui donnent des informations sur le sens impliqué. Les graphes construits ne fournissent pas l'ensemble complet des traits de chaque nœud. Lorsque un nœud virtuel est associé à un seul trait (et se trouve, par conséquent, en haut du graphe), ce trait est affiché. Il est ainsi possible d'avoir une image des traits hérités en descendant le long du graphe.

[15] Les hyponymes contiennent les traits hérités de leurs hyperonymes ainsi que les traits qui leur sont propres.

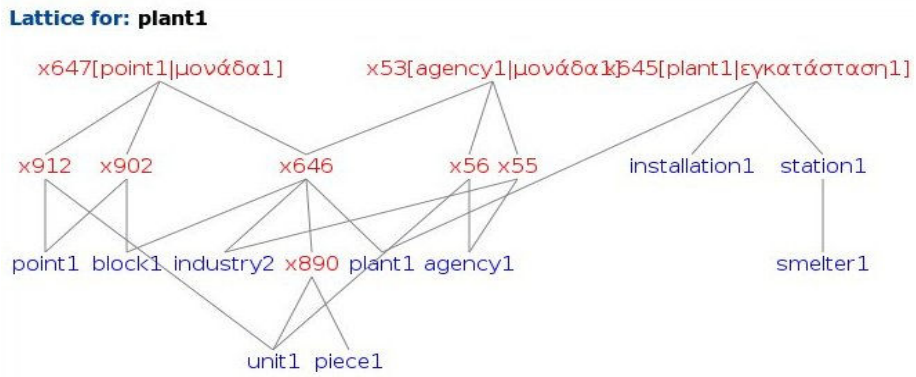


Figure 3. Le semi-treillis décrivant le sens industriel de *plant*

1.3.4. Création d'entrées de thesaurus

Les informations contenues dans les semi-treillis de traits peuvent être exploitées pour dériver des entrées similaires aux entrées d'un **thésaurus**[16]. L'entrée construite pour un mot contient des **synonymes**, des **hyperonymes,** des **hyponymes** et des **sous-sens** de chacun de ses sens (*s*). Ces relations d'un sens *s* sont définies par un agencement approprié des sens lui étant suffisamment liés dans le semi-treillis :

- un sens lié à *s*, **synonyme** ou **mot lié** (*related word*), est un sens qui partage des traits (et des traductions) avec *s*. Un synonyme est plus proche à *s* qu'un mot lié. Ces relations sont considérées comme symétriques.
- un **hyperonyme** est un sens dont *s* a hérité un trait, à la condition que le nombre des sens ayant hérité le trait en question excède un certain seuil (*SynsetLimit*[17]) ; cette condition assure que les hyperonymes possèdent des sens suffisamment génériques.

[16] Les réseaux de connexions entre les sens lexicaux décrits dans les semi-treillis sont parfois très complexes. Pour la création du thesaurus, il faut faire abstraction des détails apparaissant dans les semi-treillis.

[17] Le nombre de synonymes et de mots liés est d'autant plus grand et, à l'inverse, celui des hyperonymes et des hyponymes est d'autant plus petit, que ce seuil est élevé (la hiérarchie est plus « aplatie »).
Ce seuil peut être défini par l'utilisateur sur l'interface Web des Miroirs (http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html). Si aucune valeur n'est définie, le seuil est calculé automatiquement en fonction de la taille des champs sémantiques.

- un **hyponyme** de *s* est un sens ayant hérité un trait propre de *s*, satisfaisant la même condition à propos du nombre de sens ayant hérité le trait en question.

Un sens *s* peut également être divisé en **sous-sens** (ou nuances de sens) liés entre eux. En effet, chaque trait de *s* décrit potentiellement un sous-sens distinct. La granularité de la division en sous-sens au sein de l'entrée du thesaurus est définie par la valeur du paramètre *OverlapThreshold*[18] (seuil de recouvrement). Cette valeur peut être définie par l'utilisateur sur l'interface Web des Miroirs. La possibilité de définir ce paramètre, qui permet de modifier le nombre et la granularité des sens, est en accord avec l'hypothèse de la non existence d'une solution unique à propos de la division de la sémantique des mots en sens et en sous-sens (Kilgarriff, 1997c). Ainsi, ce paramètre permet d'adapter la granularité des sens selon les besoins.

La subdivision d'un sens en sous-sens se fonde sur l'estimation de la distinctivité des traits qui le caractérisent : si ces traits se différencient suffisamment entre eux, ils sont considérés comme des sous-sens différents. La distinctivité des traits est mesurée en termes de degré de chevauchement des ensembles de sens auxquels ils sont attribués ; cet ensemble de sens est appelé **dénotation** d'un trait. Si les dénotations des traits présentent un fort degré de recouvrement, les traits ne sont pas considérés comme représentant des sous-sens distincts, et ainsi la division en sous-sens n'a pas lieu. Au contraire, la division se justifie d'autant plus que les dénotations des traits se chevauchent moins. L'augmentation de la valeur de ce paramètre provoque alors la division d'un sens en sous-sens. Ainsi, le nombre de traits regroupés en sous-sens est d'autant plus grand que le *OverlapThreshold* est bas, tandis que le nombre de sous-sens est d'autant plus grand que le *OverlapThreshold* est élevé.

Les entrées de thesaurus créées pour certains mots polysémiques étudiés ici seront présentées dans le paragraphe suivant. Les informations sémantiques fournies au sein de ces entrées seront comparées aux résultats obtenus par notre méthode de repérage de sens.

[18] La valeur de ce paramètre doit être un nombre compris entre 0 et 1 (la valeur par défaut étant 0,05).

1.4. Comparaison de descriptions sémantiques engendrées par les deux méthodes

1.4.1. Principes sous-jacents à la comparaison

Afin que la comparaison des résultats des deux méthodes soit cohérente, nous avons adopté un ensemble de principes. Ces principes nous permettent de définir et d'appliquer une stratégie uniforme à l'égard de cette comparaison et de l'estimation de la possibilité d'utiliser les résultats des Miroirs pour la validation des résultats de notre méthode. Nous avons également envisagé la question de la complémentarité des résultats des deux méthodes et la possibilité d'enrichissement des représentations sémantiques engendrées. Les principes adoptés sont les suivants :

1. La proposition, par la méthode des Miroirs, d'une relation entre EQVs repérée au sein des clusters sert à sa validation. Les relations décrites dans les clusters sont celles repérées soit dans les deux types de contexte (anglais et grecs), soit uniquement dans les contextes anglais.
2. La proposition, par la méthode des Miroirs, d'une relation entre EQVs repérée seulement dans les contextes grecs valide la relation en question. Ces relations ne sont pas prises en compte pour la construction des clusters. Néanmoins, elles sont parfois impliquées lors du regroupement de clusters se chevauchant : deux EQVs entretenant une relation dans les contextes grecs (non présente dans les clusters de granularité fine), peuvent être liés dans un cluster de granularité grossière. La validation de ces relations assure davantage de fiabilité quant à la possibilité de regroupement des clusters se chevauchant.

1.4.2. Analyse de cas précis

Dans ce paragraphe, nous allons analyser les descriptions sémantiques fournies par la méthode des Miroirs pour les deux mots ambigus utilisés pour illustrer le fonctionnement de notre méthode d'induction de sens : *plant* et *movement*. Nous donnons en Annexe F4 le thesaurus anglais construit à partir de l'ensemble du lexique manuellement généré[19].

L'ensemble des résultats de l'application de la méthode des Miroirs sur nos données traductionnelles est consultable sur le site Web des Miroirs Sémantiques et peuvent être explorés via l'interface **Mirrors-Web**.

1.4.2.1. Comparaison des résultats obtenus pour plant

- Description sémantique de *plant* par les Miroirs

L'entrée du thesaurus créée pour le mot *plant*, par le processus présenté précédemment, est décrite dans la figure 4.

plant
Sense 1
Subsense (i)
(*Translation:* εγκατάσταση)
Synonyms: installation, smelter, station‹1›.
Subsense (ii)
(*Translation:* μονάδα)
Synonyms: agency, block‹1›, industry‹2›, piece, point, unit.
Sense 2
(*Translation:* φυτό)
Sense 3
(*Translation:* εργαστήριο)

Figure 4. Entrée du thésaurus pour *plant*

[19] Une distinction a été établie lors de la construction des thesaurus, relativement à la prise en compte des équivalents traduisant une seule fois un mot dans le corpus (les *hapax*). Même si la prise en compte de ces équivalents permet la construction d'entrées de thesaurus plus riches, elle introduit, en même temps, du bruit dans les résultats. Le thesaurus fourni en Annexe a été construit en ne prenant pas en compte ces traductions. L'autre thesaurus, qui a été construit par prise en compte de ces traductions, est consultable sur le site Web des Miroirs Sémantiques.

D'après les informations contenues dans cette entrée, *plant* a trois sens. Le premier correspond au sens industriel du mot et est divisé en deux sous-sens, dont le premier est illustré par l'EQV *εγκατάσταση* (egkatastasi) et l'ensemble de synonymes : *installation*, *smelter*, *station<1>*, et le second par l'EQV *μονάδα* (monada) et l'ensemble de synonymes : *agency*, *block<1>*, *industry<2>*, *piece*, *point*, *unit*. Les ensembles de synonymes comprennent les **sens synonymes** à un sens du mot (et non des mots synonymes). Par exemple, le mot *station* a trois sens dans le thesaurus, dont seul le premier (*station*<1>) est considéré comme synonyme du premier sous-sens du sens industriel de *plant*. En revanche, *installation* et *smelter* ont un seul sens dans le thesaurus. Les relations de synonymie au niveau des sens montrent que les mots ne sont pas des **synonymes exacts** mais plutôt des **quasi-synonymes**.

L'établissement de relations de synonymie au niveau des mots (lorsque les mots ont un seul sens dans le thesaurus) n'exclut pas la possibilité que les mots soient polysémiques. Il se peut simplement que leurs autres sens ne se manifestent pas dans le corpus et qu'ils ne soient donc pas repérés. Une telle relation au niveau des mots existe, par exemple, entre les mots *φυτό* (fyto) et *φυτάριο* (fytario). Néanmoins, si les sens métaphoriques de *φυτό*, comme « élève boutonneux » et « personne dans le coma » étaient présents dans le corpus, seul l'un de ces sens (le sens botanique) aurait été considéré comme synonyme de *φυτάριο*.

Parmi les sens illustrant le second sous-sens de *plant*, seul *industry<2>* est un synonyme pertinent. Les autres relations sont créées sur la base de traits liés à d'autres sens du mot *μονάδα*, qui sont attribués de façon erronée à *plant*. Il est intéressant de noter que le premier sens correspond bien au sens *plant2* que nous avons repéré par le regroupement des mots grecs dont les *t*-images inverses se recouvrent : *εγκατάσταση* (egkatastasi), *σταθμός* (stathmos), *μονάδα* (monada), *εργοστάσιο* (ergostasio) (cf. figure 2, §1.3.1.). Cependant, deux de ces quatre mots (*εργοστάσιο* et *σταθμός*) n'apparaissent pas dans l'entrée de *plant* ; ils sont néanmoins fournis comme synonymes du mot *εγκατάσταση*, dans l'entrée correspondante. De telles lacunes au niveau des entrées sont observées lorsque le nombre des traductions est élevé. Dans ce cas, les traductions ne sont en général pas toutes citées dans les entrées mais quelques-unes sont simplement

sélectionnées pour indiquer le sens en question. Plus précisément, les traductions fournies pour un sens sont celles qui ont contribué à la construction de ses traits. Les traductions absentes sont celles qui héritent certains traits, sans que leurs propres traits soient hérités. Il s'agit ici d'un choix d'affichage, qui n'a pas d'importance théorique.

Le deuxième sens repéré correspond au sens botanique de *plant*, qui est décrit uniquement par son EQV *φυτό*. Ce sens n'entretient pas de relation à des sens synonymes. Cette absence de relation de synonymie se manifeste dans d'autres cas et s'explique par la disponibilité d'informations traductionnelles pour une partie seulement des mots du corpus[20], qui rend compte du fait qu'il n'y a pas suffisamment d'informations sur d'autres mots, éventuellement liés aux mots étudiés.

Le troisième sens de *plant* décrit le sens d'« atelier » et est également caractérisé par un seul EQV : *εργαστήριο*.

Une remarque s'impose : l'entrée de *plant* ne comprend pas de sens distinct correspondant au mot grec *φυτάριο*. *Φυτάριο* a un seul EQV dans le corpus (*plant*), ce qui est également le cas pour le mot *φυτό*. L'intersection de leurs *t*-images inverses contient donc un seul élément (le mot *plant*), ce qui ne suffit pas à leur regroupement. D'après ce critère, un sens distinct devrait être proposé pour *φυτάριο*. La raison pour laquelle la proposition d'un tel sens n'est pas faite est que *φυτό* et *φυτάριο* partagent toutes leurs traductions (dans ce cas, une seule). Deux mots qui partagent la totalité de leurs traductions sont initialement classifiés comme des **synonymes absolus** et sont, par la suite, traités comme un seul sens. Ainsi, le trait qu'ils partagent est le [plant2|φυτό1-*] (qui correspond à la concaténation des noms de plusieurs sens) ; le nom complet du trait serait [plant2|φυτό1-φυτάριο], dans la mesure où le premier sens de *φυτό* et celui du mot *φυτάριο* sont considérés comme étant identiques. Cependant, cette relation de synonymie est repérée au sein des entrées correspondantes à *φυτό* et à *φυτάριο*.

[20] Il s'agit des informations obtenues suite aux trois inversions de la direction de traduction.

- Nos distinctions sémantiques pour *plant*

Les résultats obtenus pour *plant* par notre méthode d'induction de sens ont été analysés en détail au paragraphe 3.1.3. du chapitre 6. Nous reprenons ici les descriptions des sens obtenus à l'aide des clusters.

a. plant – {*φυτό*}
b. plant – {*μονάδα – εγκατάσταση*}
c. plant – {*μονάδα – σταθμός*}
d. plant – {*σταθμός – εργοστάσιο*}
e. plant – {*εργαστήριο*}

Figure 5. Clusters de sens proposés par notre méthode pour *plant*

Le programme SEMCLU forme cinq clusters d'EQVs, qui correspondent à des sens de *plant*. Il s'agit des sens obtenus avant l'élimination des EQVs rares ; le cluster contenant l'EQV *εργαστήριο*, ayant une fréquence inférieure à 10 dans le sous-corpus du mot *plant*, a été ensuite éliminé. Des sens de granularité plus grossière peuvent être proposés en prenant en compte les recouvrements des clusters.

a. *plant* – {*φυτό*}
b. *plant* – {{*μονάδα*, *σταθμός*}, {*μονάδα*, *εγκατάσταση*}, {*σταθμός*, *εργοστάσιο*}}
c. *plant* – {*εργαστήριο*}

Figure 6. Sens de granularité grossière de *plant*

Ces distinctions sémantiques coïncident avec les distinctions proposées par les Miroirs pour ce mot. Nous retrouvons ici la distinction principale entre le sens botanique de *plant* (illustré par le cluster (*a*)) et le sens industriel (illustré par les clusters (*b*), (*c*) et (*d*) (figure 5), regroupés en (*b*) (figure 6)), ainsi que le sens illustré par l'EQV *εργαστήριο* (*e*), qui correspond à une installation plus petite, destinée à des travaux d'un certain type.

Outre les distinctions sémantiques effectuées au niveau du mot ambigu, la comparaison des résultats des deux méthodes concerne aussi les relations sémantiques repérées entre ses EQVs. Nous avons exploré les similarités de ces

relations ainsi que la possibilité d'analyser la nature des relations des EQVs clustérisés, en considérant les résultats des Miroirs.

- Relations entre les EQVs de *plant*

Les entrées créées pour les deux EQVs du cluster (*b*) : *εγκατάσταση* (egkatastasi) et *μονάδα* (monada) sont présentées dans la figure 7.

εγκατάσταση

Sense 1

Hyperonyms: κατάρτιση‹1›

Subsense (i)

(*Translation:* plant)

Synonyms: **εργοστάσιο‹1›**, **σταθμός‹2›**

Subsense (ii)

(*Translation:* facility, establishment)

Synonyms: ίδρυμα, όργανο‹1›, όχημα‹2›, βοήθημα‹1›, διευκόλυνση‹1›, δυνατότητα‹1›, εξοπλισμός‹1›, ευκολία‹1›, κέντρο‹1›, μέσο‹1›, μηχανισμός‹1›, **μονάδα‹1›**, πρόγραμμα, σύστημα‹1›, συσκευή‹1›, υποδομή‹1›, χώρος‹1›

Subsense (iii)

(*Translation:* allocation)

Synonyms: παροχή‹1›

Related words: απόδοση‹2›, διάθεση‹2›, επίδομα‹1›, καταμερισμός‹1›, κατανομή‹1›, κατεύθυνση‹2›, κονδύλιο‹1›, παραχώρηση‹1›, τοποθέτηση‹1›

Subsense (iv)

(*Translation:* settlement)

Synonyms: κέντρο‹1›

Related words: αντιμετώπιση‹2›, διακανονισμός‹2›, διευθέτηση, εκδίκαση‹2›, επίλυση‹1›, ικανοποίηση‹1›, οικισμός‹2›, ρύθμιση‹1›

Sense 2

(*Translation:* investment)

μονάδα

Sense 1

Hyperonyms: κατάρτιση‹1›

Subsense (i)

(*Translation:* point)

Synonyms: παρατήρηση‹2›, φάση‹1›

Subsense (ii)

(*Translation:* establishment, facility, service, agency)

Synonyms: ίδρυμα, όργανο‹1›, όχημα‹2›, βοήθημα‹1›, γραφείο‹1›, διευκόλυνση‹1›, δυνατότητα‹1›, **εγκατάσταση‹1›**, εξοπλισμός‹1›, εταιρεία‹1›, ευκολία‹1›, κέντρο‹1›, μέσο‹1›, μηχανισμός‹1›, οργανισμός‹1›, πρόγραμμα, σύστημα‹1›, συσκευή‹1›, υπηρεσία‹1›, υποδομή‹1›, χώρος‹1›

Related words: αγαθό‹4›, ανάγκη, αποστολή‹1›, αρχή‹2›, δύναμη‹1›, εξυπηρέτηση‹1›, επίδοση‹1›, επιχείρηση‹1›, θητεία‹2›, κοινοποίηση‹1›, υποχρέωση, φορέας‹2›

Subsense (iii)

(*Translation:* site)

Synonyms: σημείο‹1›

Related words: έδρα‹2›, έκταση‹1›, διεύθυνση‹1›, εστία‹1›, κέντρο‹1›, περιοχή‹1›

Sense 2

(*Translation:* credit)

Figure 7. Entrées du thesaurus pour *εγκατάσταση* et *μονάδα*

La relation de similarité sémantique entre ces EQVs est révélée par l'inclusion d'un de leurs sens dans l'ensemble des sens synonymes illustrant l'un des sens de l'autre : le sens *μονάδα<1>* est inclus parmi les synonymes du deuxième sous-sens du sens *εγκατάσταση<1>*, qui est inclus à son tour parmi les synonymes du deuxième sous-sens du sens *μονάδα<1>*. Ces deux sous-sens sont décrits par les traductions *facility* (installation, usine) et *establishment* (établissement).

Les entrées générées pour les EQVs du cluster (d), *σταθμός* et *εργοστάσιο*, sont décrites dans la figure 8.

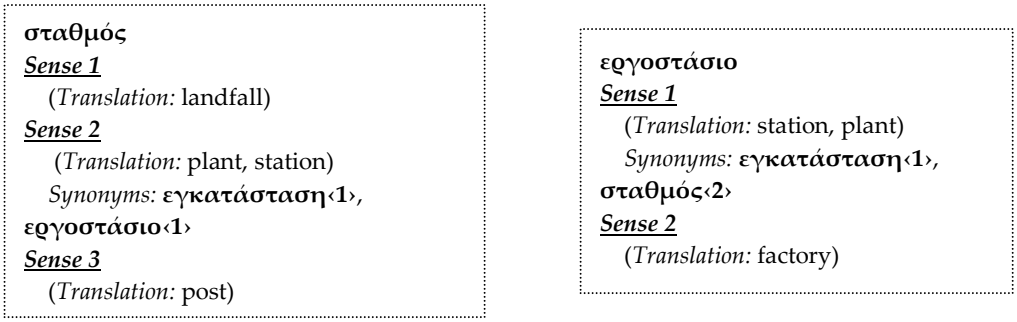


Figure 8. Entrées du thesaurus pour *σταθμός* et *εργοστάσιο*

Au sein du deuxième sens de *σταθμός*, qui correspond à son sens industriel, nous retrouvons ses relations de synonymie avec *εγκατάσταση<1>* et *εργοστάσιο<1>*. Cette relation apparaît également au sein du premier sens de *εργοστάσιο*, considéré comme synonyme de *εγκατάσταση<1>* et de *σταθμός<2>*.

Les relations *εγκατάσταση–εργοστάσιο* et *εγκατάσταση–σταθμός* n'ont pas été repérées par la méthode contextuelle. En revanche, la relation *σταθμός–εγκατάσταση*, repérée seulement dans les contextes grecs, n'a pas été considérée comme pertinente ni prise en compte par la méthode de clustering. Lors de la modification de la granularité des clusters, ces trois relations sont néanmoins trouvées au sein du cluster (b), décrivant le sens industriel de *plant* et généré par le regroupement des clusters (b), (c) et (d).

Le repérage de ces relations par les Miroirs constitue un indice de la pertinence du regroupement effectué, dans le sens où les clusters fusionnés étaient effectivement proches. Le repérage des relations en question renforce

ainsi l'hypothèse que ces relations n'avaient pas été initialement repérées en raison du manque d'informations contextuelles suffisantes.

L'équivalent *φυτό* a uniquement un sens dans le corpus, illustré par son synonyme *φυτάριο*, qui est aussi l'un des équivalents de *plant*.

φυτό
(*Translation:* plant)
Synonyms: φυτάριο

Figure 9. Entrée du thesaurus pour *φυτό*

L'absence de relations de similarité sémantique pour l'EQV *εργαστήριο*, observée lors du clustering, apparaît également dans l'entrée correspondante du thesaurus.

εργαστήριο
Sense 1
(*Translation:* laboratory)
Sense 2
(*Translation:* **plant**)
Sense 3
(*Translation:* workshop)

Figure 10. Entrée du thesaurus pour *εργαστήριο*

1.4.2.2. Comparaison des résultats obtenus pour movement

- Distinctions sémantiques de *movement* par les Miroirs

Dans la figure 11, nous présentons l'entrée du thesaurus créée pour *movement*.

movement
Sense 1
(*Translation* : **κυκλοφορία**)
Synonyms: circulation, marketing<1>, mobility<1>, transfer<2>.
Sense 2
(*Translation* : **κίνημα**)
Sense 3
(*Translation* : τάση)

Figure 11. Entrée du thesaurus pour *movement*

Le premier sens fourni dans cette entrée correspond au sens de « mouvement physique », le deuxième au sens abstrait du mot (en tant qu'organisation, association, etc.) et le dernier au sens de « tendance ». La distinction entre les deux premiers sens est repérée par notre méthode, contrairement au sens de « tendance ». L'équivalent *τάση* n'a pas été pris en compte pour le clustering, en raison de sa très basse fréquence dans le corpus.

- Nos distinctions sémantiques pour *movement*

Nous rappelons ici les sens fournis par notre méthode pour *movement*.

a. *movement* – {*μετακίνηση*, *κίνηση*, *διακίνηση*}
b. *movement* – {*κίνηση*, *διακίνηση*, *κυκλοφορία*}
c. *movement* – {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}
d. *movement* – {*κίνημα*}

Si le regroupement des clusters se chevauchant est effectué, les sens de *movement* sont décrits par les clusters suivants :

a. *movement* – {{*μετακίνηση*, *κίνηση*, *διακίνηση*}, {*κίνηση*, *διακίνηση*, *κυκλοφορία*}, {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}}
b. *movement* – {*κίνημα*}

- Relations entre les EQVs de *movement*

Les relations de similarité entre les EQVs, repérées par les Miroirs, sont décrites par les entrées qui leur correspondent (figure 12).

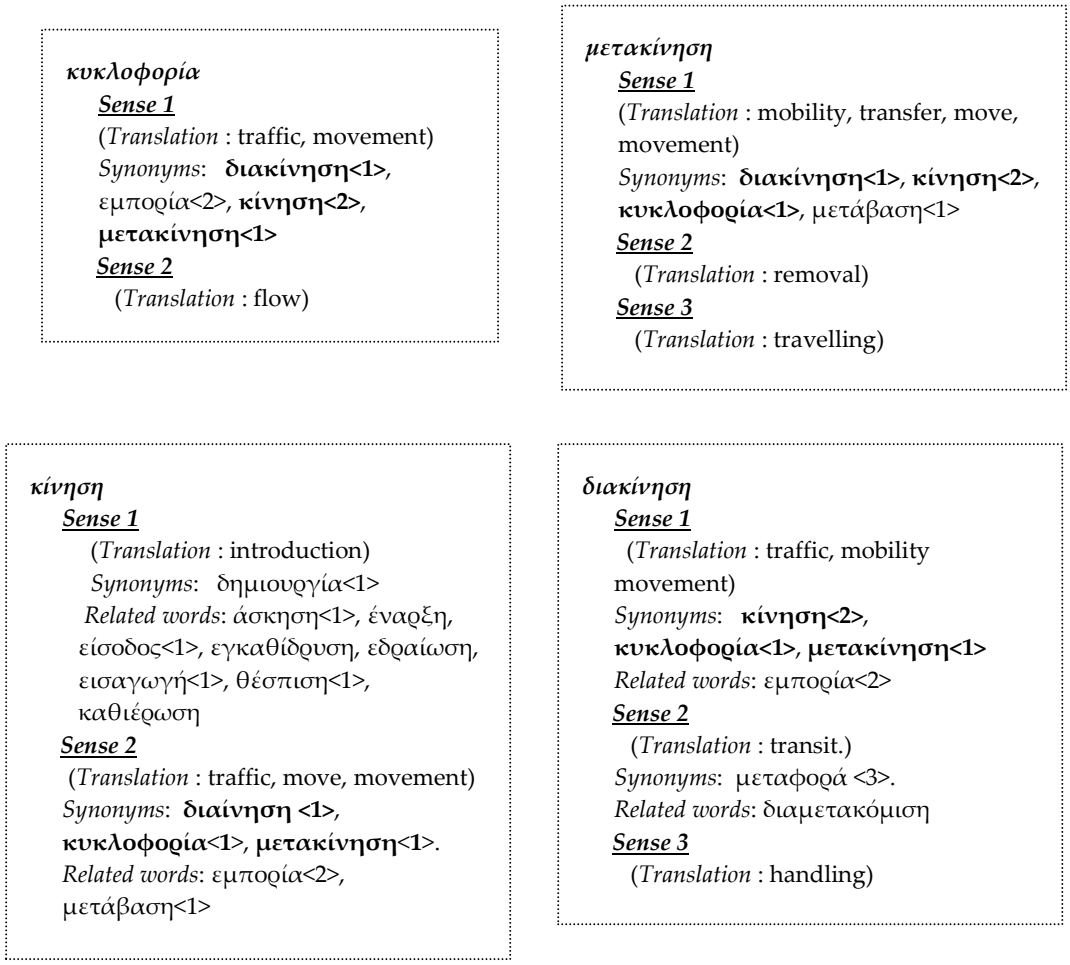


Figure 12. Entrées du thesaurus pour les EQVs de *movement*

Les relations entre les EQVs *κυκλοφορία*, *μετακίνηση*, *κίνηση* et *διακίνηση* sont trouvées au sein de ces entrées du thesaurus. Ainsi, le premier sens de *κυκλοφορία* est considéré comme synonyme du premier sens de *διακίνηση*, du second sens de *κίνηση* et du premier sens de *μετακίνηση* ; ces sens correspondent tous au sens de « mouvement physique », comme cela est rendu évident par leurs traductions (*traffic*, *mobility*, *transfer*, *move*). Les deux premières relations de *κυκλοφορία* (*κυκλοφορία – κίνηση* et *κυκλοφορία – διακίνηση*) sont trouvées dans les clusters de granularité fine, tandis que la dernière (*κυκλοφορία–μετακίνηση*) est retrouvée dans le cluster plus large (*a*), qui résulte du regroupement des clusters initialement créés.

Le repérage de cette relation par la méthode des Miroirs apporte donc plus de fiabilité pour le regroupement des clusters se chevauchant.

1.4.3. Bilan

1.4.3.1. Analyse sémantique des équivalents

L'analyse sémantique des EQVs des mots ambigus effectuée par la méthode des Miroirs donne une image de leur polysémie. Pour la plupart, il s'agit de mots polysémiques et seule une partie de leur champ sémantique est mise en relation avec le mot ambigu. Cette constatation émane de l'analyse des entrées de thesaurus créées pour les EQVs, où les relations de synonymie décrites concernent certains sens des mots grecs. Le repérage de ces relations au niveau des sens, et non à celui des mots, montre que les EQVs couvrent des champs sémantiques différents, dont seule une partie se chevauche.

Notre méthode directionnelle (anglais→grec) limite l'analyse des EQVs à leurs sens qui sont liés à ceux du mot ambigu qu'ils traduisent et qui se manifestent, par conséquent, dans son sous-corpus. L'application de notre méthode, en inversant la direction de traduction (grec→anglais), permettrait d'analyser l'ensemble de la sémantique des mots grecs à l'aide de leurs EQVs anglais dans l'ensemble du corpus d'entraînement.

1.4.3.2. Similarité des relations sémantiques proposées

De manière générale, les relations sémantiques repérées dans les entrées du thesaurus généré par les Miroirs ont une forte similarité avec celles décrites par les clusters de sens fournis par notre méthode. Certaines relations, non trouvées dans les clusters de granularité fine sont repérées dans les clusters de granularité grossière, qui résultent du regroupement de clusters se chevauchant. Cette fusion a lieu malgré le non repérage de certaines relations qui, soit ne sont détectées dans aucun type de contexte, soit le sont uniquement dans les contextes grecs et qui, ainsi, ne sont pas prises en compte lors du clustering initial.

Le regroupement des clusters se chevauchant s'appuie en effet sur l'hypothèse que le non repérage de certaines relations est dû à l'impact de la dispersion des données sur les résultats du calcul de similarité distributionnelle. Cependant, un tel regroupement implique certains risques, plus ou moins grands selon les cas, liés à la possibilité de regrouper des clusters décrivant des sens distincts (cf. §2.6.4, chapitre 6). La proposition des relations en question par la

méthode des Miroirs sert donc à valider le regroupement, en consolidant les liens entre les clusters qui se chevauchent.

1.4.3.3. Qualité des distinctions et des relations sémantiques proposées par les Miroirs

Il faut pourtant noter que, parmi les relations proposées par les Miroirs, il en existe certaines qui ne sont pas pertinentes, tant au niveau des synonymes qu'à celui des hyperonymes (par ex. la proposition de *κατάρτιση<1>* (élaboration, formation) comme hyperonyme du premier sens de *μονάδα*, qui inclut le sous-sens d'« établissement »). Notre but n'étant pas de valider la totalité des résultats de la méthode des Miroirs mais d'examiner les distinctions sémantiques proposées pour les mots ambigus étudiés et les relations proposées pour leurs EQVs, nous avons décidé d'ignorer les informations concernant des mots grecs qui ne sont pas pris en compte par notre méthode pour l'analyse du mot ambigu (par exemple, la relation de *μονάδα* avec *κατάρτιση*). Ainsi, la proposition d'une relation par les Miroirs, sans évidence contextuelle relative, n'est pas suffisante pour l'adoption de la relation. La prise en compte des informations distributionnelles pourrait éventuellement aider à la validation des résultats des Miroirs.

En outre, certaines distinctions sémantiques opérées pour un mot ne paraissent pas justifiées. Le regroupement des sens proposés en des sens plus larges serait préférable mais n'a pas lieu en raison du manque d'informations traductionnelles nécessaires (Dyvik, 2005). Il s'agit d'un problème général auquel se heurte la méthode des Miroirs, lié à l'insuffisance des traductions utilisées dans le corpus, pour les mots décrivant les sens en question, pour ce regroupement. Le problème devient plus aigu lorsque l'analyse porte sur un ensemble bien précis de mots du corpus, ce qui est le cas dans notre étude. Dans ce cas, il est possible que des traductions pouvant servir à ce regroupement soient présentes dans le corpus et qu'elles ne soient pas repérées.

1.4.3.4. Résultats de la comparaison

De façon générale, les relations sémantiques proposées par les Miroirs pour les mots grecs sont très proches de celles repérées par notre méthode. Cette similarité augmente la fiabilité de nos résultats, en consolidant ceux pour lesquels il existe peu d'évidence contextuelle. Les résultats de cette méthode, dont le fonctionnement diffère fortement de la nôtre, servent donc à valider :

- les relations entre EQVs trouvés dans les clusters générés par notre méthode
- l'hypothèse de regroupement des clusters se chevauchant, qui se base sur une généralisation des relations de similarité entre les EQVs qu'ils contiennent, dans les cas où certains d'entre eux ne sont pas explicitement liés.

2. Deuxième étape de l'évaluation qualitative

La deuxième étape d'évaluation qualitative des descriptions sémantiques fournies par notre méthode d'induction de sens consiste à comparer ces descriptions à celles issues d'une ressource lexico-sémantique multilingue prédéfinie, le réseau BalkaNet.

2.1. Le réseau BalkaNet

2.1.1. Caractéristiques de la ressource

L'objectif du projet **BalkaNet** était de construire des ressources lexico-sémantiques pour des langues qui en étaient jusqu'alors peu dotées (Stamou *et al.*, 2002 ; Tufiş *et al.*, 2004a). Plus précisément, ce projet s'appliquait à élaborer des ressources de type WordNet (Miller *et al.*, 1990) pour six langues de l'Europe du Sud-Est : grec, bulgare, roumain, turc, serbe et tchèque[21]. La structure des wordnets monolingues développés pour ces langues est identique à celle du wordnet anglais.

[21] Il s'agit d'une extension du WordNet tchèque développé dans le cadre d'EuroWordNet.

Chaque wordnet comprend des concepts organisés en taxonomies sémantiques, mis ensuite en correspondance avec leurs équivalents sémantiques dans les autres langues via un **Index Interlingue** (*Interlingual Index* (ILI)), constitué de concepts du WordNet de Princeton (version PWN 1.7. et 2.0.)[22]. ILI connecte ainsi les langues entre elles et rend possible la transition des concepts d'une langue aux concepts leur étant sémantiquement liés dans une autre. Les synsets de WordNet jouent donc le rôle de **concepts indépendants de la langue** dans cet index, qui inclut également des concepts spécifiques aux langues particulières impliquées[23]. Les réseaux individuels sont ainsi organisés en une base de données multilingue, permettant leur interconnexion.

Le développement de ces wordnets repose sur deux modèles (Tufiş *et al*., *ibid*.) :

- le **modèle d'expansion** (*expand model*) : approche basée sur les traductions, où les mots inclus dans chaque synset de PWN sont traduits de la manière la plus fidèle possible. Les relations d'un synset traduit sont, pour la plupart, importées automatiquement. Dans cette approche, de nouveaux mots peuvent être insérés dans le synset cible, tandis que des mots du synset source, difficiles à traduire, sont ignorés. Le développement est accéléré par l'utilisation d'un dictionnaire électronique bilingue.
- le **modèle de fusion** (*merge model*) : approche qui présuppose la disponibilité de ressources linguistiques structurées monolingues sur support électronique (thesaurus de granularité comparable à PWN ou dictionnaires à partir desquels une structure similaire à wordnet peut être dérivée). Le format de ces ressources est transformé pour être compatible avec celui de WordNet et les sens sont liés aux concepts de l'Index Interlingue. Les noms des relations du wordnet cible sont les mêmes que ceux des relations de PWN, même si leur topologie diffère.

La ressource de BalkaNet s'appuie sur une combinaison de ces deux modèles afin de bénéficier de leurs avantages : le modèle d'expansion assure la

[22] Cet alignement a également été effectué au sein d'EuroWordNet (Vossen, 1999).
[23] Par exemple, des synsets décrivant des nourritures typiques d'un pays, concepts non lexicalisés en anglais.

comparabilité des wordnets monolingues, tandis que le modèle de fusion garantit l'inclusion de **propriétés propres** aux langues particulières.

Les informations du réseau BalkaNet sont accessibles via l'outil **Cilix**, sous la forme d'une représentation graphique[24]. Cet outil permet la navigation entre les concepts et les termes des wordnets et fournit, en outre, des informations sur leur sémantique, les ontologies conceptuelles auxquelles ils appartiennent ainsi que les relations sémantiques les liant à d'autres termes et concepts trouvés dans les wordnets.

2.1.2. Applications envisagées pour BalkaNet

L'application finale envisagée pour cette ressource est la **Recherche d'Information Multilingue** (RIM) (Tufiş *et al*., 2004a)[25]. La logique de l'utilisation de BalkaNet dans des tâches de recherche d'information consiste à avancer qu'une représentation conceptuelle structurée du domaine d'intérêt, liée aux wordnets multilingues, peut aider les utilisateurs à repérer des informations de manière précise en utilisant dans les requêtes des mots clés de leur propre langue. L'ontologie de BalkaNet permet d'indexer des documents du Web sur la base de leur **similarité conceptuelle**. La précision de l'alignement interlingue et la grande couverture des wordnets particuliers sont essentielles à la performance de cette application.

L'ontologie conceptuelle peut également faciliter la prise en compte de **paraphrases**, en établissant des connexions entre les termes utilisés dans une requête et des termes leur étant sémantiquement liés, éventuellement repérés dans les documents indexés (Stamou *et al*., 2004). Dans cette perspective, l'ontologie peut alors être considérée comme un « guide » vers une organisation plus signifiante des sources de données indexées par les moteurs de recherche. L'approche de l'indexation conceptuelle permet ainsi d'effectuer de la RIM **basée sur le contenu** (c'est-à-dire, ne nécessitant pas de correspondances exactes avec les mots clés) (Ambroziak et Woods, 1998). De manière générale, l'exploitation de

[24] Au moment de cette étude, l'outil Cilix était disponible à l'URL suivante : http://ru2146.cti.gr/Cilix/CilixSetup.exe

[25] Pour l'application de EuroWordNet dans la RIM, voir Peters *et al*. (1998), Vossen *et al*. (1999),Verdejo et al. (2000).

ressources sémantiques multilingues dans un tel cadre est considérée augmenter la **précision** des résultats (par la distinction des sens lexicaux) ainsi que le **rappel** (par l'identification de termes conceptuellement liés) (Gonzalo *et al*., 2000).

Néanmoins, les distinctions sémantiques présentes au sein de ce type de ressources ne s'avèrent pas toujours appropriées à ce cadre, dans la mesure où des sens distingués ne conduisent pas systématiquement à des topics ou à des types de documents différents et ne facilitent pas, par conséquent, la RIM (Gonzalo *et al*., *ibid*.). L'exploitation de ces ressources dans ce but est donc souvent associée à l'utilisation de techniques de clustering, permettant de diminuer la granularité des sens décrits (Peters *et al*. 1998). Ces techniques présentent, néanmoins, d'autres inconvénients, étant donné qu'elles ne conduisent pas toujours à des clusters de mots sémantiquement similaires (Peters *et al*., *ibid*. ; Gonzalo *et al*., *ibid*.).

Dans un certain nombre de travaux, BalkaNet a également été exploité à des fins de **désambiguïsation lexicale** (Ion et Tufiş, 2004 ; Tufiş *et al*., 2004b,c). Cette application présente pourtant des inconvénients similaires à ceux observés lors de l'exploitation de WordNet pour la WSD (cf. §2.3.3, chapitre 3), qui concernent :

- la faible couverture de la ressource
- l'arbitraire des sens décrits, reflétée dans
  - l'absence d'informations liées à des domaines spécialisés
  - l'inclusion d'informations concernant des usages très rares des mots
- la granularité très fine des distinctions sémantiques établies.

L'utilisation de BalkaNet dans le cadre de la Traduction n'a pas, quant à elle, été envisagée. Nous estimons que la structure des données au sein de cette ressource rend difficile son exploitation dans des applications traductionnelles. Les méthodes de WSD exploitant BalkaNet qui ont été proposées (Ion et Tufiş, 2004 ; Tufiş *et al*., 2004b; c) nécessitent en effet que la traduction du mot source soit connue. Plus précisément, ces méthodes opèrent sur des corpus parallèles et le processus de WSD consiste à extraire, pour chaque mot d'une paire de mots

alignés, les codes ILI des synsets les contenant. Ce processus fournit deux listes de codes ILI, une pour chaque langue, et la WSD est effectuée en identifiant, soit le code qui se trouve à l'intersection des deux listes, soit la paire de codes, parmi ceux contenus dans les deux listes, qui correspondent aux concepts les plus similaires. La traduction du mot source doit donc être connue pour que l'intersection des codes correspondant aux mots puisse être calculée. Pourtant, dans une tâche de traduction, la traduction d'une nouvelle occurrence d'un mot polysémique n'est pas connue au départ.

Cependant, si une méthode de WSD opératoire dans un cadre de traduction et utilisant les informations de BalkaNet pouvait être proposée, le sens identifié pour le mot source serait représenté par un ensemble de synonymes, sauf dans le cas où un synset contiendrait un seul EQV dans la LC. Si le synset sélectionné pour décrire le sens d'une instance était constitué de plus d'un EQV, la sélection de l'EQV le plus adéquat pour traduire la nouvelle instance du mot source n'irait pas de soi, sauf si la méthode de WSD permettait la considération d'informations venant des nouveaux contextes ou qu'une méthode de sélection lexicale était utilisée.

2.2. Comparaison de nos résultats aux descriptions sémantiques fournies par BalkaNet

Nous avons procédé à une étude des descriptions sémantiques fournies par BalkaNet pour les mots ambigus *plant* et *movement* et pour leurs EQVs de traduction en grec. Le but était de comparer ces descriptions aux résultats obtenus par notre méthode.

2.2.1. Descriptions sémantiques de *plant*

Nous présentons tout d'abord les sens décrits pour *plant* dans BalkaNet et mis en correspondance avec des synsets grecs. Les sens du mot sont décrits, dans la figure 13, par les synsets correspondants de WordNet et par leur définition (*gloss*). Entre parenthèses, nous traduisons en français la définition de chaque sens.

1. {**flora, plant, plant life**} : a living organism lacking the power of locomotion (*organisme vivant dépourvu de capacité de mouvement*)
2. {**industrial plant, plant, works**} : buildings for carrying on industrial labor (*bâtiments consacrés aux activités industrielles*)
3. {**plant**} : something planted secretly for discovery by another (*ce qui est caché secrètement (par quelqu'un) pour être découvert par quelqu'un d'autre*)
4. {**plant**} : an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience (*acteur situé dans le public et dont le jeu a été travaillé mais paraît improvisé au public*)

Figure 13. Sens du nom *plant* dans le ILI

Les deux premiers sens de *plant* correspondent à ses sens homonymiques, le sens botanique et le sens industriel, qui ont été également repérés par notre méthode. En revanche, les sens 3 et 4 sont très rares. Des instances du mot véhiculant ces sens n'apparaissent pas dans notre corpus d'apprentissage. Le fait que les sens en question ne soient pas relatifs aux domaines des textes traités explique pourquoi notre méthode ne fournit pas d'informations à leur égard.

La proposition de sens non relatifs aux domaines traités constitue, en effet, un inconvénient de l'exploitation de ressources prédéfinies pour la WSD. La considération de ces sens augmente le nombre de choix possibles lors de la désambiguïsation ; le système de WSD est alors, en effet, contraint d'opérer une sélection parmi tous les sens proposés – ce qui complique le traitement sans bénéfice évident. Ces deux caractéristiques des inventaires de sens prédéfinis, à savoir la rareté de certains sens et, fréquemment, l'absence de sens spécifiques aux domaines traités, sont considérées comme des inconvénients de ces ressources (Pantel et Lin, 2002).

Les synsets illustrant les sens rares de *plant* (3 et 4) ne contiennent que le mot *plant* (et non un synonyme) et ne retiennent aucune relation (hyponymie, méronymie, etc.) avec d'autres synsets au sein du réseau. En revanche, les synsets décrivant les sens homonymiques du mot (1 et 2) sont liés à d'autres synsets. Leurs relations directes sont décrites dans la figure 14. Des descriptions plus complètes de ces parties du réseau, incluant les relations héritées par les synsets, sont incluses en Annexe F1.

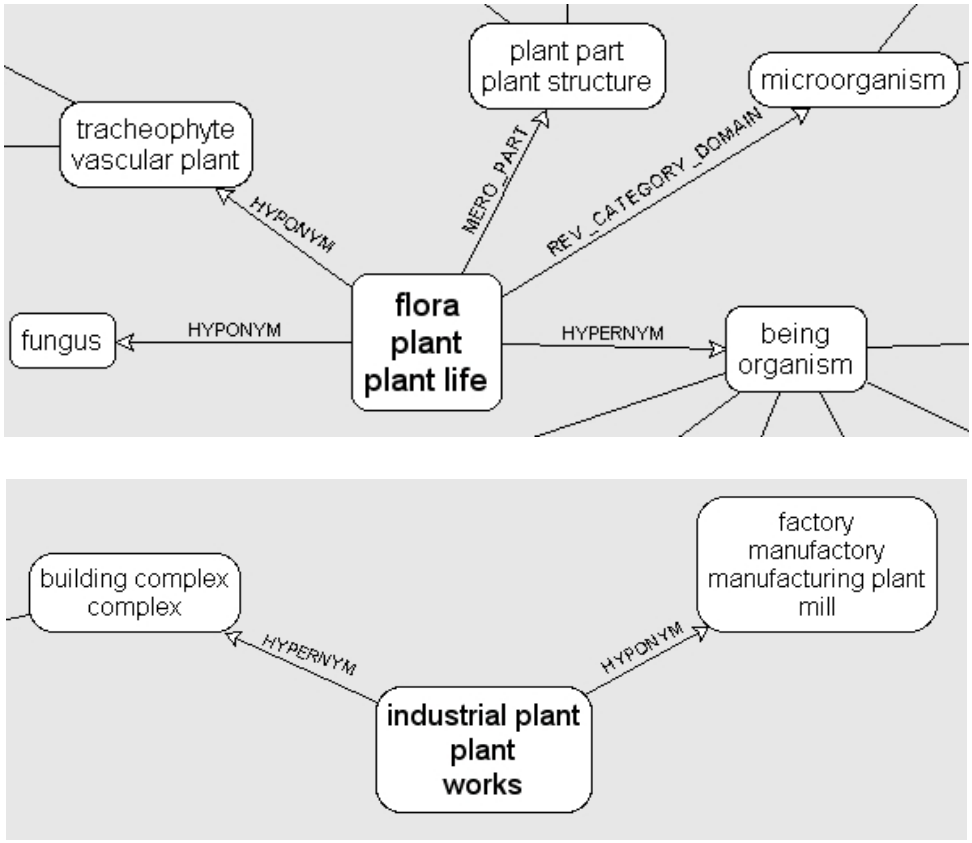


Figure 14. Synsets décrivant les sens industriel et botanique de *plant* au sein du réseau

Dans la mesure où le wordnet anglais constitue l'ILI, les sens du mot *plant* sont liés à des synsets indiquant les mêmes sens en grec (figure 15). Les définitions (*glosses*) des sens grecs sont identiques aux définitions anglaises (cf. figure 13). Entre parenthèses, nous traduisons en français les mots grecs trouvés dans les synsets.

1. **φυτό** (fyto) (*plante)*
2. **βιομηχανικές εγκαταστάσεις** (viomixanikes egkatastaseis) (*installations industrielles*)
3. **παγίδα** (pagida) (*piège*)
4. **ηθοποιός** (ithopoios) (*acteur*)

Figure 15. Synsets du wordnet grec correspondant aux sens de *plant*

Les informations contenues dans les synsets grecs sont relativement limitées, chaque synset contenant un seul mot ou un seul terme composé : le

synset illustrant le premier sens de *plant*, le sens botanique, contient l'EQV *φυτό*, et celui décrivant le sens industriel contient un terme composé de l'EQV *εγκατάσταση* (« installation ») et de l'adjectif *βιομηχανικός* (« industrielle »). Les clusters fournis par notre méthode et illustrant ces deux sens contiennent des informations plus riches. Nous y retrouvons, d'une part, les deux EQVs *φυτό* et *εγκατάσταση*, qui constituent les deux premiers synsets. Mais d'autre part, le cluster de *εγκατάσταση* contient un autre mot grec qui lui est sémantiquement similaire : l'EQV *μονάδα*. En regroupant les clusters se chevauchant, l'ensemble des mots sémantiquement liés s'enrichit encore des EQVs *σταθμός* et *εργοστάσιο*.

Les relations du synset grec décrivant le sens botanique de *plant* sont décrites dans la figure 16. L'hyperonyme du synset du mot *φυτό* (*plant*) est le synset contenant le mot *οργανισμός* (organismos, « organisme ») ; son hyponyme est le synset du mot *μανιτάρι* (manitari, « champignon ») ; le synset de *φυτό* entre également dans une relation de méronymie avec le synset du mot *ίνα* (ina, « fibre »). Ces relations nous paraissent très pauvres : de nombreux hyponymes et méronymes auraient pu en effet être ajoutés à ce réseau. Nous n'insisterons néanmoins pas sur ce point, puisque seules les informations trouvées au sein des synsets décrivant les sens des mots nous intéressent. Nous présentons ici visuellement leurs relations afin de faciliter la compréhension des sens qu'ils décrivent.

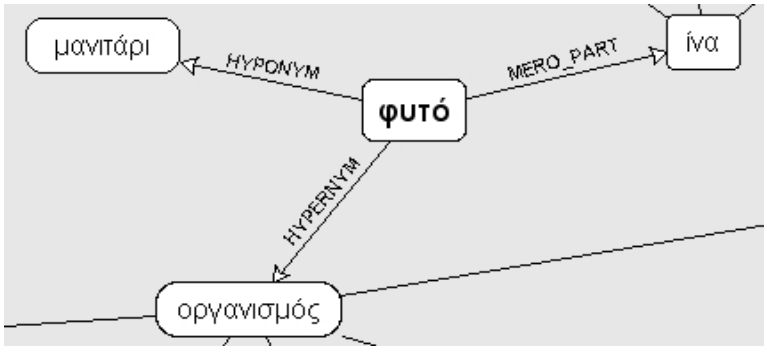


Figure 16. Relations du synset grec décrivant le sens botanique de *plant*

Les relations du synset grec décrivant le sens industriel de *plant* sont illustrées dans la figure 17. Le synset *βιομηχανικές εγκαταστάσεις*

(viomichanikes egkatastaseis, « installations industrielles ») est l'hyperonyme du synset contenant *εργοστάσιο* (ergostasio, « usine »), tandis que son propre hyperonyme est le synset *κτηριακό συγκρότημα* (ktiriako sygkrotima, « ensemble de bâtiments »).

Figure 17. Relations du synset grec décrivant le sens industriel de *plant*

Les synsets grecs correspondant aux deux sens rares de *plant* sont constitués des mots grecs *παγίδα* (pagida) (sens 3), qui signifie « piège », et *ηθοποιός* (ithopoios) (sens 4), qui signifie « acteur ». Ces mots pourraient difficilement être considérés comme des traductions de *plant* en grec. Les informations fournies par les synsets en question sont très pauvres : chaque synset ne contient qu'un seul mot et n'entretient pas de relations avec d'autres synsets du réseau.

Les divergences observées à propos de la quantité d'informations fournies pour les différents sens se justifient par la différence de statut des sens en question. Néanmoins, aucune information n'est fournie au niveau des entrées de la ressource sur la différence de statut entre les sens homonymiques de *plant* (1 et 2) et ses sens plus rares (3 et 4). Cette difficulté des lexiques d'énumération de sens à distinguer le **statut des sens** (Dolan, 1994 ; Pustejovsky, 1995 : 29) a déjà été soulignée[26]. Les sens décrits au sein de ces lexiques se situent tous au même niveau et se voient donc attribués le même statut, ce qui ne permet pas une différenciation de traitement lors d'un processus de WSD.

En revanche, notre méthode d'acquisition de sens permet la prise en compte des relations inter-sens. Les **sens mutuellement exclusifs** sont décrits par

[26] Voir paragraphe 2.3.3.2 du chapitre 3, sur les avantages et les inconvénients liés à l'exploitation de ressources préétablies.

des clusters distincts, tandis que les **sens** et les **sous-sens liés** sont décrits par des clusters présentant des recouvrements. Cette distinction s'avère utile tant au niveau du traitement, qu'au niveau de l'évaluation des résultats d'une méthode de WSD, où une pénalisation différenciée des erreurs de désambiguïsation (Resnik et Yarowsky, 1997, 2000) serait possible.

2.2.2. Descriptions sémantiques de *movement*

Les sens proposés dans le wordnet anglais pour *movement* sont les suivants :

1. {**motion, move, movement**} : the act of changing location from one place to another (*action de changer de lieu, d'aller d'un endroit à un autre*)
2. {**motility, motion, move, movement**} : a change of position that does not entail a change of location (*changement de position qui n'entraîne pas un changement de lieu*)
3. {**motion, movement**} : a natural event that involves a change in the position or location of something (*événement naturel qui implique un changement de position ou de lieu de quelque chose*)
4. {**front, movement, social movement**} : a group of people with a common ideology who try together to achieve certain general goals (*groupe de personnes ayant une idéologie commune, qui essaient ensemble d'atteindre certains buts généraux*)

Figure 18. Sens du nom *movement* dans le ILI

Les trois premiers sens de *movement* sont de granularité très fine et pourraient être regroupés au sein d'un sens de granularité plus grossière, celui de « mouvement physique ». La distinction entre ces sens n'est pas évidente et le serait d'autant moins pour un programme informatique qui devrait en sélectionner un, lors d'une tâche de désambiguïsation. En outre, la sélection par un programme de WSD d'un sens parmi des sens très proches pourrait provoquer une perte d'informations (Dolan, 1994). Cette finesse des descriptions sémantiques constitue, comme nous l'avons déjà souligné, l'une des critiques les plus importantes formulées dans le cadre des travaux de WSD et d'étiquetage sémantique exploitant WordNet (cf. §2.3.3, chapitre 3).

Ces trois sens proches se situent au même niveau que le sens 4, qui se distingue plus clairement et correspond au sens abstrait du mot (le sens de « mouvement social »). L'attribution du même statut aux sens mutuellement exclusifs et aux sens liés ne permet pas leur traitement différencié. Cette non

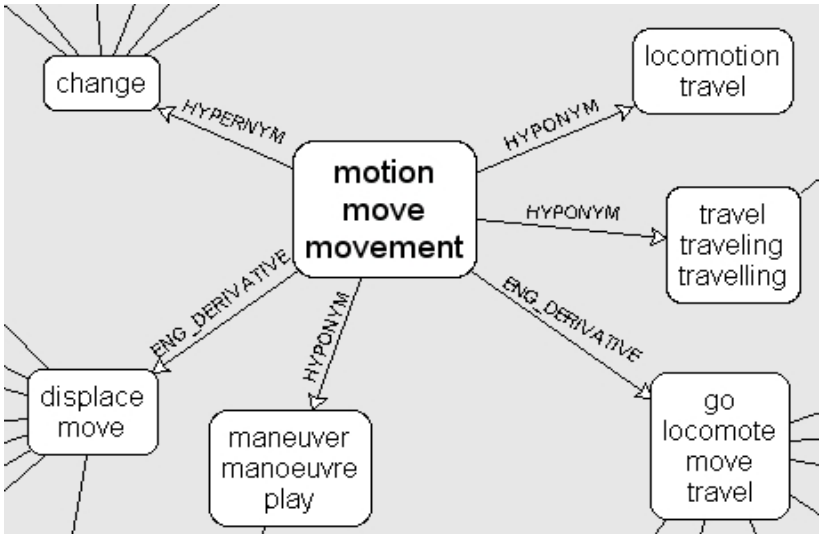
distinction des sens relativement à leur statut aurait également un impact négatif lors de l'évaluation d'un processus de WSD, dans le sens où des erreurs concernant des sens proches et distincts seraient pénalisées de la même manière (Resnik et Yarowsky, 1997, 2000). Afin d'éviter cela, il faudrait utiliser des méthodes permettant de découvrir les liens entre les sens, qui rendraient possible leur clustering[27]. Ces méthodes pourraient consister, par exemple, en l'exploitation de la similarité des mots inclus dans les synsets ; en la considération des relations sémantiques décrites au sein de WordNet (Peters *et al*., 1998 ; Mihalcea et Moldovan, 2001) ou en la prise en compte des similarités au niveau des définitions fournies pour les sens en question dans WordNet ou dans d'autres ressources (Navigli, 2006 ; Navigli *et al*., 2007). Par exemple, deux synsets ayant le même hyperonyme pourraient être considérés comme sémantiquement liés ; de même pour les synsets dont les définitions contiennent les mêmes mots (comme c'est le cas, ici, pour les définitions des trois premiers sens de *movement* qui contiennent toutes les mots : *change*, *location*, *position*).

Les relations entre les synsets des sens en question et d'autres synsets du réseau, illustrées dans la figure 19, permettent en effet de capter certaines de leurs similarités[28]. Les synsets décrivant les deux premiers sens ont certaines relations en commun (le même hyperonyme : *change*, et la même relation dérivationnelle : *move*). En revanche, la relation d'hypéronymie qui caractérise le synset du quatrième sens (figure 20) le distingue clairement des autres. Ces relations inter-sens ne sont pourtant pas définies au sein de la ressource ; le recours à des méthodes permettant leur repérage à partir des informations fournies dans le réseau s'avère donc nécessaire.

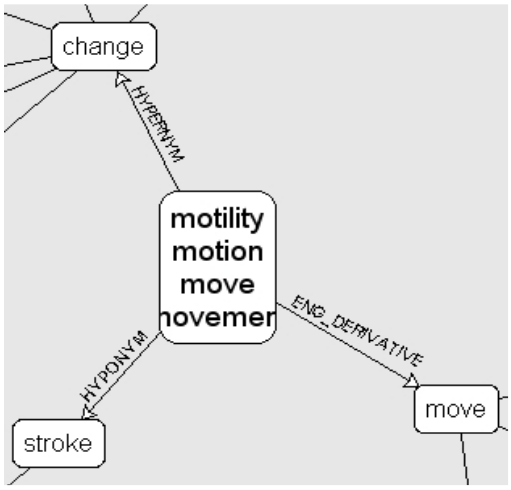
[27] Pour une description détaillée des méthodes qui ont été proposées dans ce but, cf. §2.3.3.3, chapitre 3.

[28] La relation 'eng_derivative', décrite dans la figure 19, capte une relation de dérivation entre un nom et un verbe en anglais.

1.



2.



3.

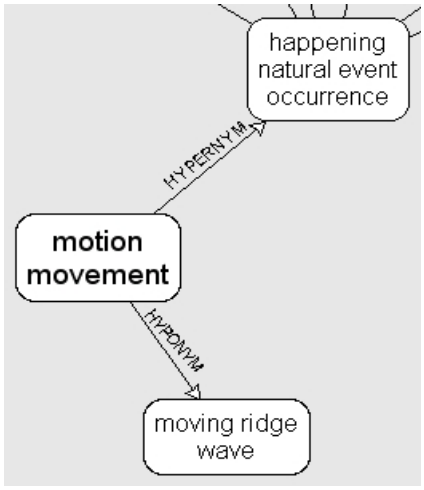


Figure 19. Relations des synsets décrivant les trois premiers sens de *movement*

4.

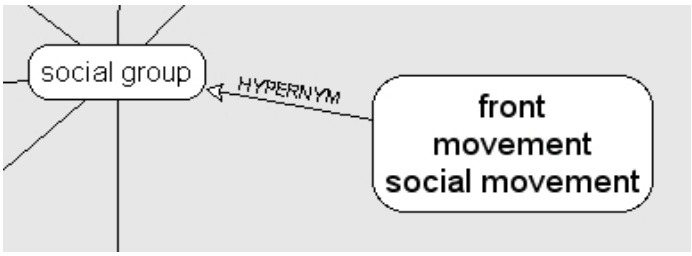


Figure 20. Relations du synset décrivant le quatrième sens de *movement*

Les synsets grecs correspondant à ces sens ne clarifient pas non plus les distinctions sémantiques proposées, ni *a fortiori* celle établie entre le premier et le deuxième sens du mot.

1. αλλαγή θέσης (alagi thesis) (*changement de position*)
2. αλλαγή θέσης (alagi thesis) (*changement de position*)
3. **κίνηση** (kinisi) (*motion*)
4. **κίνημα** (kinima) (*mouvement social*)

Figure 21. Synsets grecs correspondant aux sens de *movement*

Le troisième sens est décrit par le mot grec *κίνηση*, contenu aussi dans deux des clusters d'EQVs obtenus par notre méthode (clusters *a* et *b*).

a. *movement* – {*μετακίνηση*, ***κίνηση***, *διακίνηση*}
b. *movement* – {***κίνηση***, *διακίνηση*, *κυκλοφορία*}
c. *movement* – {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}
d. *movement* – {*κίνημα*}

Tout comme c'était le cas pour *plant*, les informations repérées dans les clusters de sens de *movement* sont plus riches que les informations fournies au sein du wordnet. Ainsi, le mot grec *κίνηση* est lié aux mots *μετακίνηση*, *διακίνηση* et *κυκλοφορία*, qui lui sont sémantiquement liés. Si le regroupement des clusters sur la base de leurs recouvrements a lieu, *κίνηση* sera également lié à l'EQV *κινητικότητα*, qui lui est aussi sémantiquement proche. En regroupant les clusters, nous obtenons alors les deux sens suivants :

a. le sens de **mouvement physique** : *movement* – {{*μετακίνηση*, *κίνηση*, *διακίνηση*}, {*κίνηση*, *διακίνηση*, *κυκλοφορία*}, {*μετακίνηση*, *διακίνηση*, *κινητικότητα*}}
b. le sens de **mouvement abstrait** : *movement* – {*κίνημα*}

La possibilité de regrouper des clusters se chevauchant permet donc de modifier la granularité des sens proposés. Les deux sens qui en résultent correspondent aux sens de granularité grossière du mot auxquels seraient réduits

les sens de WordNet si une diminution de leur granularité était souhaitée et donc que les trois sens de granularité trop fine étaient regroupés. Le regroupement des clusters peut donc être conçu comme remplissant la même fonction que les méthodes d' « ambiguïsation » adoptées pour la réduction de la granularité des sens proposés dans WordNet (Dolan, 1994 ; Mihalcea et Moldovan, 2001).

CONCLUSION

Dans ce chapitre, nous avons comparé les résultats obtenus pour deux mots ambigus (*plant* et *movement*) par notre méthode d'acquisition de sens, d'une part, avec les descriptions sémantiques fournies par la méthode des Miroirs Sémantiques pour les mêmes mots et, d'autre part, avec les descriptions sémantiques de ces mots fournies par une ressource lexico-sémantique prédéfinie.

La comparaison de nos résultats à ceux des Miroirs présente l'avantage que les deux types de résultats ont été obtenus à partir des mêmes données. Ainsi, les résultats sont relatifs aux domaines représentés dans notre corpus d'apprentissage. En revanche, la comparaison de nos résultats aux descriptions sémantiques fournies par une ressource lexico-sémantique existante, met en évidence l'intérêt d'utiliser une méthode d'acquisition de sens basée sur les données, même dans le cas où de telles ressources sont disponibles.

La similarité des descriptions sémantiques obtenues par la méthode des Miroirs – qui s'appuie uniquement sur des informations traductionnelles n'exploitant pas d'informations du contexte des mots – avec les descriptions fournies par notre méthode –qui combine des informations des deux types – permet de valider nos résultats. De plus, la comparaison de ces descriptions avec les représentations sémantiques de BalkaNet montre les avantages de notre méthode par rapport à un lexique d'énumération de sens. Ces avantages sont, plus précisément, la considération des relations inter-sens et la possibilité de modifier automatiquement la granularité des sens obtenus, en fonction des besoins qui se présentent.

La prochaine étape d'évaluation, exposée dans le chapitre suivant, concernera la possibilité d'exploiter des correspondances sémantiques inter-langues établies par notre méthode dans des tâches précises, à savoir la WSD et la sélection lexicale pour la traduction.

CHAPITRE 10

EVALUATION QUANTITATIVE DES METHODES DE DESAMBIGUÏSATION ET DE SELECTION LEXICALE

INTRODUCTION

L'évaluation quantitative effectuée s'applique aux méthodes de désambiguïsation et de sélection lexicale présentées dans le chapitre 7. La méthode de désambiguïsation s'appuie sur les résultats du clustering réalisé lors de l'étape d'acquisition de sens. L'attribution d'un sens à une instance d'un mot ambigu en contexte correspond ici à la sélection d'un des clusters décrivant les différents sens du mot. La proposition effectuée par la méthode de WSD se base sur les informations provenant des contextes source de la nouvelle instance.

La méthode de sélection lexicale entreprend, quant à elle, de filtrer les résultats de la méthode de WSD, si besoin est. La nécessité d'un tel filtrage repose

sur le choix effectué par la méthode de WSD pour une instance du mot ambigu. Si un cluster de plusieurs EQVs est choisi, la méthode de sélection retient, parmi les EQVs, celui qui est le plus apte à traduire l'instance en question. En revanche, dans le cas où le cluster illustrant le sens de la nouvelle instance ne contient qu'un seul EQV, la méthode de sélection lexicale n'intervient pas. Contrairement à la méthode de WSD, la méthode de sélection lexicale s'appuie sur des informations provenant du contexte de la LC.

L'évaluation quantitative de ces deux méthodes permet de mesurer l'impact de l'exploitation des clusters de sens sur les résultats des processus de WSD et de sélection lexicale. Cet impact est évalué en comparant les résultats des deux méthodes à ceux d'une « **méthode de base** » (*baseline*), méthode simple et rapide à implémenter que nous présenterons plus loin (cf. §2.4.).

1. Corpus d'évaluation

Le corpus parallèle utilisé pour l'évaluation (corpus d'évaluation ou corpus de test) diffère du corpus d'apprentissage. Il s'agit de la partie anglais-grec du corpus parallèle multilingue **EUROPARL** (Koehn, 2005). Les caractéristiques et les étapes de prétraitement de ce corpus ont déjà été présentées en détail (cf. §2, chapitre 4). Nous rappelons simplement ici que dans la première version du corpus (1.1), qui est celle que nous avons utilisée, les textes sont alignés au niveau phrastique : une **unité de traduction** est constituée d'un segment de la LS et d'un segment de la LC liés par une relation de traduction. Nous parlons ici de segments (et non de phrases) dans la mesure où un segment peut contenir plus d'une phrase (maximum deux), comme c'est le cas du segment source de l'unité de traduction présentée dans la figure 1.

> Here, it is not a question of seeking a localised cure, but rather of triggering an overall **movement** in society through the development of human resources. It is a question of investing in people rather than in infrastructures. -- *Εδώ το θέμα δεν είναι να αναζητήσουμε ένα μεμονωμένο αποτέλεσμα επούλωσης, αλλά να υποκινήσουμε μια συνολική* ***κίνηση*** *στην κοινωνία μέσω της ανάπτυξης του ανθρώπινου δυναμικού, το θέμα είναι να επενδύουμε στον άνθρωπο πιο πολύ παρά στις υποδομές.*

Figure 1. Unité de traduction dont le segment source contient deux phrases

Les méthodes de WSD et de sélection lexicale comparent des informations provenant des nouveaux contextes des mots ambigus avec celles retenues lors de l'apprentissage. Pour que cette comparaison soit possible, le corpus de test a dû être lemmatisé et étiqueté morphosyntaxiquement.

Des **échantillons de test** ont été ensuite automatiquement construits à partir de ce corpus pour chaque mot ambigu étudié (cf. §2.2.4, chapitre 4). Chaque échantillon contient les unités de traduction où le mot ambigu apparaît dans le segment de la LS. Nous avons, au total, élaboré **160 échantillons de test**, dont 10 correspondent aux mots ambigus anglais du lexique anglais-grec manuellement construit et 150 correspondent aux mots anglais du lexique anglais-grec automatiquement généré.

L'échantillon constitué pour chaque mot ambigu est trié sur la base de ses EQVs de traduction, fournis dans l'entrée correspondante au mot au sein du lexique bilingue. Chaque ensemble d'unités de traduction contient un seul EQV du mot dans le segment de texte de la LC et pourrait être désigné comme *MotAmbigu_EQV*. Par exemple, l'unité de traduction présentée dans la figure 1 appartient à l'ensemble *movement_κίνηση* tandis que celle décrite dans la figure 2 appartient à l'ensemble *movement_κυκλοφορία*.

> The free **movement** of persons is one of the basic freedoms achieved through European integration.
> *Η ελεύθερη* ***κυκλοφορία*** *προσώπων είναι μία εκ των θεμελιωδών ελευθεριών, που κατοχυρώθηκαν μέσω της ευρωπαϊκής ενοποίησης.*

Figure 2. Unité de traduction extraite de l'ensemble *movement_κυκλοφορία*

Le mot *movement* est traduit dans EUROPARL par tous les EQVs repérés au sein du corpus d'apprentissage (*κυκλοφορία* (kykloforia), *κίνηση* (kinisi), *διακίνηση* (diakinisi), *μετακίνηση* (metakinisi), *κινητικότητα* (kinitikotita), *κίνημα* (kinima)). Six ensembles d'unités de traduction ont été ainsi créés à partir de son échantillon de test, chacun correspondant à un de ses EQVs[1]. L'EQV utilisé pour traduire le mot ambigu dans chaque ensemble d'unités de traduction

est appelé la **traduction de référence**. Les unités de traduction contenant plus d'un EQV de traduction du mot ambigu sont éliminées.

Les traductions du corpus de test provenant de traducteurs professionnels, elles sont considérées comme étant de très bonne qualité.

L'évaluation est faite séparément pour chaque mot ambigu. A la fin, nous calculons des scores globaux pour l'ensemble des mots de chacun des deux lexiques.

2. Evaluation de la méthode de désambiguïsation lexicale

2.1. Estimation de la justesse des prédictions de désambiguïsation

2.1.1. Absence d'un corpus sémantiquement annoté

Dans le cas de la méthode de WSD, est évaluée sa capacité à sélectionner le sens correct, parmi tous les sens possibles, pour une nouvelle instance d'un mot ambigu. L'idéal, dans le cadre d'une telle évaluation, serait de disposer d'un corpus d'évaluation sémantiquement étiqueté par les sens de notre inventaire, et où les instances des mots seraient étiquetées par leur sens en contexte. Ce corpus constituerait un étalon d'or (*gold standard*) contenant les réponses correctes, réponses qui seraient utilisées pour l'évaluation (Kilgarriff, 1998b). En faisant appel à un corpus annoté pour l'évaluation, le résultat du processus de WSD serait considéré comme correct si le sens proposé pour une instance correspond à celui décrit par son étiquette.

Le corpus sémantiquement annoté le plus fréquemment utilisé est le corpus monolingue SemCor (Miller *et al.*, 1994 ; Landes *et al.*, 1998)[2]. Les mots de contenu de ce corpus sont manuellement annotés par les sens des mots fournis dans WordNet. Cependant, l'évaluation d'une méthode de WSD reposant sur un

[2] Un corpus parallèle (anglais-italien) sémantiquement annoté (MultiSemCor) a été créé sur la base de SemCor, par transfert des annotations sémantiques des mots d'une langue aux mots auxquels ils sont alignés dans une autre (Bentivogli *et al.*, 2004). Ainsi, les deux côtés du bitexte sont annotés par référence au même inventaire de sens, WordNet.

inventaire de sens différent de celui utilisé pour l'annotation du corpus n'est pas possible. L'emploi d'un tel corpus serait en effet possible à la condition qu'une méthode permettant la mise en correspondance (*mapping*) des sens fournis par les deux inventaires ou qu'un dictionnaire décrivant ces correspondances soient disponibles. Et, si SemCor a été utilisé pour l'entraînement de systèmes de WSD supervisés, sa petite taille limite l'extensibilité des systèmes en question.

2.1.2. Principes d'évaluation

Ne disposant donc pas d'un corpus d'évaluation sémantiquement étiqueté par les sens de notre inventaire, nous avons conçu un processus d'évaluation de la méthode de WSD qui repose sur les principes suivants :

1. le sens proposé pour une instance d'un mot ambigu est considéré comme **correct** :
   a. si un **cluster d'un EQV** est proposé et que l'EQV en question correspond à la traduction de référence
   b. si un **cluster de plusieurs EQVs** est proposé et que la traduction de référence est repérée dans le cluster proposé.
2. le sens proposé pour une instance d'un mot ambigu est considéré comme **erroné**
   a. si un **cluster d'un EQV** est proposé et que l'EQV ne correspond pas à la référence
   b. si un **cluster de plusieurs EQVs** est proposé et que la traduction de référence n'appartient pas au cluster.

2.1.3. Exploitation des traductions pour l'évaluation

La traduction d'une instance d'un mot ambigu dans le corpus parallèle de test peut être considérée comme son **étiquette sémantique**. Ce type d'étiquetage est adopté dans les tâches multilingues de la campagne Senseval. Plus précisément, dans Senseval-3 (Ckhlovski *et al*., 2004)[3], les étiquettes de sens des

[3] Le but de cette tâche est de créer un cadre pour l'évaluation des systèmes de TA avec une focalisation sur la traduction des mots ambigus.

mots ambigus anglais correspondent à leurs traductions en hindi, et, dans la tâche correspondante de SemEval (Jin *et al*., 2007), les étiquettes des mots ambigus chinois sont constituées de leurs traductions en anglais[4]. Dans ces deux tâches, les instances des mots ambigus sont manuellement étiquetées par leurs traductions dans l'autre langue. Les systèmes de WSD participant à ces tâches doivent donc prédire la traduction correcte de chaque instance des mots ambigus. Cette utilisation des traductions en tant qu'étiquettes sémantiques permet une évaluation de granularité grossière, étant donné qu'une traduction peut correspondre à plusieurs sens d'un mot ambigu (Jin *et al*., *ibid*.).

De même, dans la tâche d'échantillon lexical pour l'anglais via des textes parallèles anglais-chinois de SemEval (Chan & Ng, 2005 ; Ng *et al*., 2003 ; Ng et Chan, 2007), les étiquettes de sens des mots anglais correspondent à leurs traductions en chinois. Cette tâche est similaire à la tâche multilingue de Senseval3 (Ckhlovski *et al*., *ibid*.), à la différence que les exemples d'entraînement et de test sont collectionnés sans annoter manuellement les instances des mots ambigus. En revanche, les exemples sémantiquement étiquetés sont recueillis (semi-)automatiquement à partir de textes parallèles anglais-chinois, alignés au niveau des mots : les traductions chinoises correspondant aux différents sens de WordNet (1.7.1.) d'un mot anglais sont sélectionnées à la main et attribuées aux sens. Là aussi, une traduction peut correspondre à plusieurs sens, ce qui diminue la granularité des distinctions sémantiques présentes dans WordNet. Les données de la partie anglaise des textes parallèles servent de données d'entraînement et de test, puisqu'elles sont considérées comme désambiguïsées et sémantiquement annotées.

Dans ces tâches, ce qui est évalué est la capacité des systèmes de WSD à prédire la traduction correcte pour de nouvelles instances des mots. Dans notre méthode d'évaluation, la **traduction de référence** trouvée dans une unité de traduction peut être également conçue comme une **étiquette du sens** de l'instance correspondante du mot source, étiquette qui renvoie à un sens décrit par un cluster. Ainsi, nous évaluons, par calcul, la capacité de la méthode de WSD à prédire le sens véhiculé par l'étiquette attribuée à chaque instance de test.

[4] Le but de cette tâche est de créer un cadre pour l'évaluation de la WSD dans des systèmes de TA chinois-anglais. Le corpus chinois est d'abord sémantiquement annoté, puis chaque sens est remplacé par sa traduction en anglais, traduction pouvant regrouper des sens différents.

Néanmoins, le compteur qui calcule les prédictions correctes ne s'incrémente pas uniquement dans les cas de correspondance exacte entre proposition et référence.

Une telle correspondance n'est recherchée que dans le cas de sélection, par la méthode de WSD, d'un cluster à un EQV : dans ce cas, si l'EQV correspond à la référence, l'instance est considérée comme correctement désambiguïsée. Les clusters à un seul élément sont toujours disjoints de ceux décrivant les autres sens du mot, ce qui signifie qu'ils décrivent des sens distincts. Dans ce cas, la traduction de référence est considérée comme décrivant un sens du mot ambigu (décrit par le cluster en question), ne pouvant pas être véhiculé par d'autres EQVs.

2.1.4. Métrique de précision enrichie

Les traductions de référence trouvées dans le corpus résultent d'un processus de décision effectué par le traducteur[5]. Cela ne signifie pas, néanmoins, que la solution retenue soit la seule possible dans un contexte donné. Lors de l'évaluation, les informations sémantiques relatives à l'EQV qui constitue la référence, acquises lors de l'entraînement, peuvent donc être également prises en compte. Il s'agit des relations de similarité sémantique éventuellement entretenues entre l'EQV en question et d'autres EQVs du mot ambigu, relations modélisées dans les clusters de sens. Ainsi, dans le cas de sélection d'un cluster à plusieurs (>1) EQVs, si la traduction de référence est contenue dans le cluster, la nouvelle occurrence est alors considérée comme étant correctement désambiguïsée.

Si seule la traduction qui constitue l'étiquette d'une instance était considérée comme correcte (traduction de référence), l'évaluation pourrait être conçue comme se fondant sur le principe de **précision**, sous-jacent aux métriques d'évaluation utilisées pour la TA. Les faiblesses des métriques d'évaluation du résultat de la TA basées sur la notion de précision, ont déjà été soulignées (cf.

[5] La décision du traducteur est influencée par des paramètres assez variés : hormis le système d'instructions conscientes et inconscientes, objectives (dépendantes du matériel linguistique) et subjectives qui influence le choix d'une unité lexicale (Levý, 1967), cette décision dépend aussi de la dimension pragmatique du travail de traduction, dans le sens où elle doit assurer l'effet souhaité dans un type de texte donné.

§1.3, chapitre 8). D'un point de vue quantitatif, ce critère strict a un impact négatif sur les résultats de l'évaluation. Dans une perspective qualitative, il ne permet pas de prendre en compte le fait que des EQVs de traduction similaires peuvent correspondre au même sens d'un mot source et, ainsi, qu'un mot ambigu en contexte peut avoir plus d'une bonne traduction ; caractéristique qui constitue une propriété essentielle du processus de traduction[6].

Notre méthode d'évaluation du résultat de la WSD pourrait être considérée comme une métrique de **précision enrichie**, qui exploite les relations paradigmatiques (relations de similarité sémantique, comme la synonymie, la quasi-synonymie, etc.) repérées entre les EQVs de traduction des mots ambigus. Les résultats de notre méthode de WSD ne consistent pas en des mots isolés de la LC, mais en des clusters de traductions décrivant des sens. C'est pourquoi, l'essentiel n'est pas d'établir une proposition correspondant de manière exacte à la référence, mais bien plutôt une proposition qui décrive le sens de l'instance du mot ambigu, décrit, également, par son EQV (référence). Ce sens peut ainsi être représenté par des EQVs de traduction alternatives, sémantiquement plausibles.

2.2. Possibilité d'évaluation par exploitation de sens de granularité variable

Outre les clusters de sens initialement fournis par la méthode d'acquisition de sens, la méthode de WSD peut exploiter les clusters de granularité grossière, créés par le fusionnement de clusters se chevauchant. En effet, dans certains cas, des prédictions, pourtant sémantiquement correctes, ne sont pas prises en compte lors de la considération des clusters de granularité fine. Tel est le cas de la proposition de désambiguïsation pour l'instance suivante de *movement* :

> *I am therefore delighted that steps are finally being taken which will allow the theoretical freedom of* ***movement*** *of persons to be translated into practice, albeit still far from perfect practice!*
>
> *Επομένως είμαι ικανοποιημένος που επιτέλους λαμβάνονται μέτρα τα οποία θα επιτρέψουν να γίνει πράξη η αρχή της ελεύθερης* ***κυκλοφορίας*** *των*

[6] A ce propos, Vickrey *et al*. (2005) soutiennent qu'un module de traduction des mots dans un système de TA devrait pouvoir proposer des alternatives ordonnées en fonction de leur qualité au lieu de produire des prédictions uniques.

προσώπων, έστω και αν τα μέτρα αυτά είναι ακόμα, κατά τη γνώμη μου, ανεπαρκέστατα!

La traduction de référence de *movement* est l'EQV *κυκλοφορία* (kykloforia). Pourtant, la proposition de la méthode de WSD concerne la paire d'EQVs *μετακίνηση–κινητικότητα* (metakinisi-kinitikotita), qui renvoie au cluster {*μετακίνηση, διακίνηση, κινητικότητα*} (metakinisi-diakinisi-kinitikotita), seul cluster contenant la paire d'EQVs en question. La traduction de référence n'étant pas repérée dans ce cluster, l'instance est considérée comme étant erronément désambiguïsée (principe 2(b)).

Cette prédiction aurait dû, néanmoins, être considérée comme correcte. Le sens illustré par le cluster proposé est celui de « mouvement physique », qui est effectivement celui véhiculé par cette instance de *movement* ; deux des trois EQVs du cluster (*μετακίνηση* et *διακίνηση*) pourraient aisément traduire l'instance en question, à la place de l'EQV *κυκλοφορία*. Si le troisième EQV (*κινητικότητα*) est sémantiquement proche des autres, il constitue néanmoins dans ce contexte une traduction moins adéquate (véhiculant plutôt le sens de « mobilité »). Cependant, l'important lors de l'évaluation de la méthode de WSD est le sens décrit par le cluster sélectionné ; la sélection de l'EQV le plus adéquat en contexte serait, quant à elle, opérée ensuite, par la méthode de sélection lexicale.

Etant donné que des prédictions sémantiquement correctes, comme celle décrite ci-dessus, ne peuvent être prises en compte sur la base des clusters de granularité fine, il serait possible d'évaluer les résultats de la méthode de WSD en exploitant les sens de granularité grossière. Ces sens correspondraient aux clusters élargis, générés par le regroupement des clusters se chevauchant. Ainsi, dans le cas de notre exemple, la prédiction de la paire d'EQVs *μετακίνηση–κινητικότητα* (metakinisi-kinitikotita) serait alors considérée comme correcte, car appartenant au même cluster de granularité grossière que la traduction de référence *κυκλοφορία* :

{{***μετακίνηση***, *κίνηση, διακίνηση*}, {*κίνηση, διακίνηση,* ***κυκλοφορία***},
{***μετακίνηση***, *διακίνηση,* ***κινητικότητα***}}

Ce cluster illustre le sens de « mouvement physique » par fusion des clusters plus petits décrivant ce sens et les distingue du cluster véhiculant le sens de « mouvement abstrait » {*κίνημα*} (kinima).

Le découpage de l'évaluation des méthodes de WSD en deux étapes en fonction de la granularité des sens distingués se retrouve dans Senseval-3 (Mihalcea *et al*., 2004). Parmi les données fournies aux participants, il existe une **carte sémantique** (*semantic map*) qui regroupe les sens de WordNet en des sens plus grossiers. Ce découpage se justifie par la faible performance des systèmes de WSD lorsque le nombre de sens fourni dans l'inventaire est très élevé et leur distinction difficile.

Nous considérons que nos distinctions sémantiques ne sont pas aussi fines que celles opérées dans WordNet. C'est pour cette raison que nous évaluons la méthode de WSD n'exploitant que les clusters initialement générés. Nous souhaitons ici seulement signaler qu'il serait possible de tirer profit de la « flexibilité » de notre inventaire de sens et de procéder à une WSD utilisant des sens plus grossiers. En général, l'évaluation des méthodes de WSD exploitant de telles distinctions donne des résultats plus élevés que celle des méthodes exploitant des sens fins (Ng et Chan, 2007).

2.3. Mesures utilisées pour l'évaluation

Les mesures utilisées pour l'évaluation quantitative des résultats de notre méthode de WSD sont les taux de **rappel** et de **précision**:

$$Rappel = \frac{nombre\ de\ prédictions\ correctes}{nombre\ de\ nouvelles\ instances}$$

$$Précision = \frac{nombre\ de\ prédictions\ correctes}{nombre\ de\ prédictions\ \ faites\ par\ le\ système}$$

Le taux de rappel indique la proportion des nouvelles instances d'un mot ambigu qui sont correctement désambiguïsées, tandis que la précision indique,

quant à elle, la proportion des prédictions de désambiguïsation faites par le système qui sont correctes.

Nos résultats sont ensuite comparés aux résultats issus d'une méthode de base (*baseline*). Les résultats de cette méthode correspondent aux deux scores, précision et rappel, comme nous allons l'expliquer dans le paragraphe suivant. Pour faciliter la comparaison entre nos résultats et la baseline, nous avons également eu recours à la **f-mesure** (van Rijsbergen, 1979), score combinant précision et rappel en une mesure unique. Il s'agit, plus précisément, de la **moyenne harmonique** du rappel et de la précision, calculée selon la formule suivante :

$$f\text{ - }mesure = \frac{2 * (précision * rappel)}{précision + rappel}$$

2.4. Méthode de base

La méthode de base la plus couramment employée dans les évaluations Senseval consiste à sélectionner **le sens le plus fréquent** d'un mot ambigu pour toutes ses instances (Miller *et al*., 2004). Ce sens, défini sur la base des informations fournies pour le mot dans l'inventaire de sens (WordNet), correspond au premier sens fourni pour le mot dans l'inventaire.

Notre méthode de WSD exploite les sens décrits par les EQVs de traduction des mots ambigus. La méthode de base (*baseline*) adoptée ici consiste alors à proposer **l'EQV le plus fréquent** d'un mot ambigu comme illustrant le sens de toutes ses nouvelles instances. Cet EQV est celui qui traduit le plus fréquemment le mot dans le corpus d'apprentissage.

L'heuristique du sens le plus fréquent, bien que pouvant être considérée comme simpliste, s'avère **très puissante**, étant donné l'asymétrie de la distribution des sens des mots dans des textes réels, autrement dit le fait que certains mots soient beaucoup plus souvent utilisés dans un sens que dans un autre. Cette asymétrie se reflète également au niveau des EQVs utilisés pour traduire les différents sens d'un mot source dans la LC. Comme nous le montrerons en effet dans le paragraphe suivant, l'EQV le plus fréquent d'un mot

dans le corpus d'apprentissage est le plus souvent celui qui traduit le plus grand nombre d'instances du mot dans le corpus de test.

Nous comparons au score de la baseline les scores de rappel, de précision et de f-mesure obtenus. Le **score de la baseline** correspond tant au **rappel** qu'à la **précision**. Ce score est calculé sur la base du nombre d'instances pour lesquelles le sens proposé (illustré par l'EQV le plus fréquent) est correct ; ce nombre coïncide ainsi avec la fréquence d'occurrence de l'EQV le plus fréquent dans l'échantillon de test du mot ambigu.

Par exemple, dans le cas du mot *movement*, où l'EQV le plus fréquent est *κυκλοφορία*[7], le nombre d'instances correctement désambiguïsées en proposant cet EQV est égal à sa fréquence dans l'échantillon de test du mot (1461). Pour calculer le rappel, ce nombre, qui correspond aux prédictions correctes, est divisé par le nombre total des nouvelles instances de *movement* (1461/2622), tandis que, pour calculer la précision, ce nombre est divisé par le nombre de prédictions faites par le système (1461/2622). Le nombre total des nouvelles instances et le nombre de prédictions coïncident, étant donné que l'EQV le plus fréquent est proposé pour toutes les instances. Les deux scores sont ainsi identiques et correspondent au score de la baseline (0,057).

2.5. Résultats de l'évaluation pour les mots du lexique bilingue manuellement généré

La première étape de l'évaluation de notre méthode de WSD concerne les mots du lexique bilingue manuellement généré. Dans le Tableau 1, nous donnons le nombre d'instances des mots anglais de ce lexique trouvées dans leurs échantillons de test, ainsi que le nombre d'instances de chaque mot traduites par chacun de ses EQVs. Le mot grec donné en gras correspond à l'EQV le plus fréquent du mot dans le corpus d'apprentissage, EQV servant à calculer la *baseline*. Le nombre total des instances de test des mots ambigus est : 14.456.

[7] La fréquence de cet EQV dans le corpus d'apprentissage est de 318, tandis que le deuxième EQV le plus fréquent du mot (*διακίνηση*) apparaît seulement 39 fois.

Mot ambigu (fréq)	EQVs (fréq.)
competence (908)	**αρμοδιότητα (763)**
	ικανότητα (87)
	δικαιοδοσία (32)
	επάρκεια (14)
	δυνατότητα (7)
	κατάρτιση (4)
	δεξιότητα(1)
movement (2192)	κυκλοφορία (1222)
	κίνημα (386)
	διακίνηση (286)
	κίνηση (174)
	μετακίνηση (89)
	κινητικότητα (35)
occupation (236)	κατοχή (158)
	επάγγελμα (42)
	δραστηριότητα (14)
	απασχόληση (12)
	κατάληψη (10)
	ασχολία (0)
plant (900)	εγκατάσταση (304)
	φυτό (232)
	σταθμός (140)
	εργοστάσιο (164)
	μονάδα (59)
	εργαστήριο (1)
	φυτάριο (0)
communication (2866)	**ανακοίνωση (2017)**
	επικοινωνία (774)
	ενημέρωση (35)
	πληροφορία (17)
	κοινοποίηση (12)
	ανταλλαγή (5)
	αναφορά (3)
	διαβίβαση (2)
	γνωστοποίηση (1)
paper (581)	έγγραφο (247)
	βίβλος (65)
	βιβλίο (23)
	εργασία (21)
	κείμενο (23)
	χαρτί (202)
	δοκιμασία (0)

Mot ambigu (fréq)	EQVs (fréq.)
preparation (952)	**προετοιμασία (759)**
	κατάρτιση (33)
	εκπόνηση (32)
	παρασκεύασμα (63)
	παρασκευή (11)
	επεξεργασία (22)
	προπαρασκευή (19)
	σκεύασμα (13)
facility (447)	**εγκατάσταση (188)**
	διευκόλυνση (85)
	δυνατότητα (86)
	μέσο (33)
	υπηρεσία (27)
	εξοπλισμός (14)
	υποδομή (10)
	ίδρυμα (4)
treatment (1200)	μεταχείριση (661)
	θεραπεία (154)
	αντιμετώπιση (139)
	επεξεργασία (107)
	περίθαλψη (54)
	αγωγή (36)
	καθεστώς (19)
	διαχείριση (18)
	εξέταση (10)
	κατεργασία(0)
	χορήγηση(2)
power (4174)	**εξουσία (1662)**
	αρμοδιότητα (914)
	ενέργεια (487)
	δύναμη (675)
	ισχύς (216)
	δυνατότητα (127)
	ικανότητα (39)
	δικαιοδοσία (22)
	καθήκον (20)
	ευχέρεια(12)

Tableau 1. Taille des échantillons de test des mots ambigus et fréquence des EQVs

Lorsque le nombre de segments de test correspondant à un EQV est égal à zéro (par ex. pour l'EQV de *treatment* : *κατεργασία* (katergasia)), cela signifie que l'EQV en question n'a pas été rencontré dans l'échantillon de test du mot ambigu. Nous rappelons aussi que les instances des mots ambigus traduites dans le corpus de test par d'autres EQVs que ceux trouvés dans le corpus d'apprentissage ne sont pas prises en compte.

Il est aussi à noter que lorsqu'un mot ambigu est contenu plusieurs fois dans une unité de traduction, une seule prédiction est faite. La raison en est que la méthode de WSD n'utilise pas d'informations sur l'ordre des mots. Ainsi, en exploitant le contexte en tant que « sac de mots », les différentes instances du mot ambigu se voient attribuées le même sens. Nous considérons donc, pour l'évaluation, qu'il n'y a qu'une seule instance de test.

Dans la Figure 3, nous présentons la f-mesure de la méthode de WSD obtenue pour chacun des mots de ce lexique, en les comparant avec la baseline.

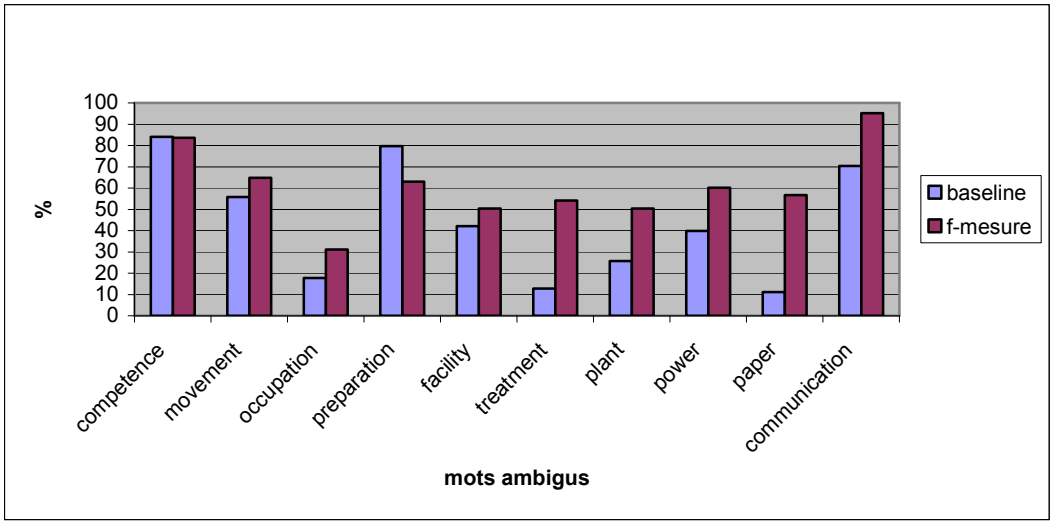


Figure 3. Evaluation de la WSD pour chaque mot du lexique manuellement généré

Les cas où les résultats de la baseline sont faibles sont ceux où l'EQV le plus fréquent dans le corpus d'apprentissage ne correspond pas à la référence la plus fréquente dans le corpus de test. Tel est le cas, par exemple, de *treatment*, où l'EQV le plus fréquent *θεραπεία* (therapeia) traduit 154 fois le mot dans l'échantillon de test correspondant et où l'EQV utilisé le plus fréquemment pour

ce mot est *μεταχείρισ*η (metacheirisi) (661 fois) (cf. Tableau 1). Dans le cas de *paper* également, l'EQV le plus fréquent *βίβλος* (vivlos) apparaît plus rarement dans l'échantillon de test du mot que *έγγραφο* (engrafo) (65 *vs* 247 fois).

D'une manière générale, la **f-mesure de notre méthode dépasse aisément le score de la baseline**. Le seul cas où la baseline est clairement supérieure est celui du mot *preparation*, où l'EQV le plus fréquent *προετοιμασία* (proetoimasia) est celui qui est largement le plus fréquent dans l'échantillon de test du mot (traduisant 79,73% des nouvelles instances). Le **score global** de la f-mesure obtenu par la méthode de WSD pour les mots du lexique manuellement généré est donné dans la Figure 4, où il est, là aussi, comparé à la performance de la méthode de base pour tous les mots de ce lexique.

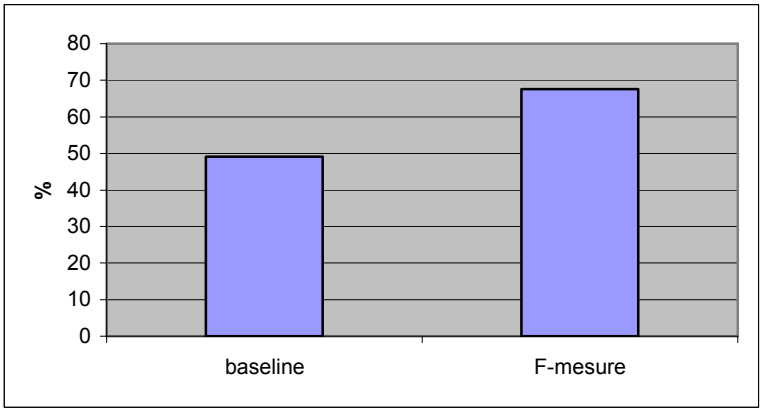


Figure 4. Evaluation de la WSD pour la totalité des mots du lexique manuellement généré

2.6. Résultats de l'évaluation pour les mots du lexique bilingue automatiquement généré

La deuxième étape de l'évaluation de notre méthode de WSD concerne les mots du lexique bilingue automatiquement généré. La taille des échantillons de test élaborés pour ces mots est donnée en Annexe D2. La taille des échantillons est calculée en nombre d'instances des mots ambigus. Nous donnons le nombre total d'instances d'un mot anglais et le nombre d'instances traduites par chacun de ses EQVs. Le mot grec donné en gras correspond à l'EQV le plus fréquent du mot dans le corpus d'apprentissage, EQV servant à calculer la *baseline*.

Le score global de la f-mesure, calculé sur la base des résultats obtenus pour la totalité des mots de ce lexique, est donné dans le Figure 5. Au sein de cette figure, il est comparé au score obtenu par la méthode de base pour la totalité des mots.

Les résultats de cette évaluation confirment la pertinence des résultats obtenus par notre méthode. Ici aussi, la f-mesure (76,99%), calculée pour la mesure de WSD, **dépasse aisément** les résultats de la méthode de base (51,42%).

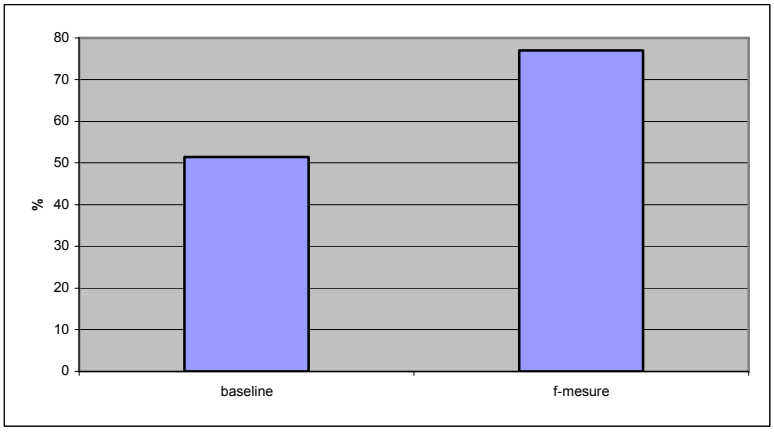


Figure 5. Evaluation de la WSD pour la totalité des mots du lexique automatiquement généré

2.7. Discussion sur les résultats de la méthode de désambiguïsation lexicale

2.7.1. Remarques sur les résultats

Les résultats de la méthode de WSD exploitant les clusters de sens sont **très encourageants**. La différence entre la f-mesure de cette méthode et le score de la méthode de base est **remarquable**, tant pour le lexique manuellement généré que pour le lexique automatiquement généré. Cette différence indique l'impact positif de l'exploitation des clusters de sens sur les résultats de l'évaluation de la méthode de WSD.

Les sens d'un mot ambigu n'étant pas décrits par chacun de ses EQVs séparément mais par des clusters d'EQVs, situant les descriptions sémantiques à

un niveau d'abstraction plus élevé, l'évaluation des sens proposés ne nécessite pas de correspondances exactes à la référence.

2.7.2. Comparaison aux résultats obtenus dans d'autres tâches d'évaluation

Nos résultats sont comparables à ceux obtenus dans la tâche d'échantillon lexical pour l'anglais (Mihalcea *et al*., 2004) et la tâche multilingue d'échantillon lexical (Chklovski *et al*., 2004)[8] de Senseval-3, ainsi que dans la tâche d'échantillon lexical chinois-anglais de SemEval (Jin *et al*., 2007). Mais, contrairement à notre évaluation, qui concerne uniquement la désambiguïsation de noms, les évaluations de Senseval s'appliquent en outre aux verbes et aux adjectifs. Néanmoins, nous estimons qu'une comparaison de nos résultats à ceux de la seule campagne d'évaluation des systèmes de désambiguïsation existante est intéressante.

L'inventaire de sens utilisé pour les noms dans la tâche d'échantillon lexical pour l'anglais de Senseval-3 (Mihalcea *et al*., *ibid*.) est WordNet (version 1.7.1.)[9]. La possibilité d'évaluer les méthodes relativement à des distinctions sémantiques plus grossières existe également, en utilisant, comme nous l'avons déjà souligné, une carte sémantique qui regroupe les sens fournis dans cet inventaire. La méthode de base utilisée dans le cadre de cette évaluation est celle qui attribue à chaque mot son sens le plus fréquent dans WordNet. Le score de la baseline (qui correspond à la précision et au rappel) est de 55,2% pour les distinctions sémantiques fines et de 64,5% pour les distinctions sémantiques grossières. La performance de la plupart des systèmes de WSD participant à cette tâche (supervisés ou non) dépasse de manière significative la baseline : le meilleur système atteint une performance de 72,9% et de 79,3%, respectivement pour les sens fins et grossiers.

La différence entre la tâche multilingue d'échantillon lexical et celle d'échantillon lexical réside en ce que l'inventaire de sens utilisé n'est pas extrait

[8] L'échantillon lexical utilisé comprend 41 mots en tout (18 noms, 15 verbes et 8 adjectifs) avec des degrés de polysémie différents, comme cela se perçoit au nombre de traductions possibles en hindi (Chklovski *et al*., 1004).

[9] WordNet constitue l'inventaire de sens utilisé pour les noms et les adjectifs. Un autre inventaire a été utilisé pour les verbes, en raison de la performance très faible des systèmes de WSD concernant cette catégorie grammaticale, probablement due au grand nombre de sens définis dans WordNet.

d'un dictionnaire mais constitué à partir de l'ensemble des traductions des mots ambigus dans une autre langue (hindi)[10]. Les contextes des mots ambigus utilisés pour la désambiguïsation sont donc en anglais, tandis que leurs « étiquettes sémantiques » correspondent à leurs traductions possibles en hindi. Les données d'entraînement et de test utilisées pour une partie des mots de l'échantillon lexical sont sémantiquement étiquetées[11].

L'ensemble des systèmes participant à cette tâche sont des systèmes supervisés. Ils donnent tous des résultats supérieurs à la baseline, ce qui est interprété par Chklovski *et al*. (*ibid*.) comme un indice du fait que les distinctions sémantiques effectuées par les traductions sont claires et fournissent les informations suffisantes à l'apprentissage de classificateurs par les méthodes supervisées. La précision et le rappel des systèmes sont identiques, dans la mesure où ils fournissent tous une réponse pour la totalité des instances de test. Un score d'exactitude est ainsi rapporté, qui mesure la proportion de décisions correctes.

Les résultats obtenus sur les **données sémantiquement annotées** dans cette tâche dépassent ceux obtenus sur les **données étiquetées uniquement par les traductions**. Dans le premier cas, le score de la baseline est de 55,8% et est inférieur au score de tous les systèmes ; l'exactitude du système réalisant la meilleure performance est de 67,3%. Dans le deuxième cas, la baseline est de 51,9 % ; tous les systèmes atteignent de meilleurs résultats que ceux de la baseline, le meilleur système ayant une exactitude de 63,4%. Cette différence indique que les informations sémantiques sont probablement utiles pour la tâche de traduction des mots ambigus.

Dans la tâche multilingue d'échantillon lexical chinois-anglais de SemEval (Jin *et al*., 2007), les deux systèmes non supervisés impliqués obtiennent des scores inférieurs au score de la baseline du sens le plus fréquent, contrairement aux systèmes supervisés, qui dépassent tous nettement la baseline.

[10] Selon la suggestion de Resnik et Yarowsky (1999).

[11] Pour la majorité des mots (31/41), les données d'entraînement et de test comprennent des instances extraites du British National Corpus (BNC), tandis que pour les 10 mots restant, les données utilisées sont extraites des données de la tâche d'échantillon lexical anglais (English lexical sample) de Senseval-2 et sont donc sémantiquement étiquetées.

3. Evaluation de la méthode de sélection lexicale

3.1. Estimation de la justesse des prédictions de sélection lexicale

La sélection lexicale consiste à sélectionner la traduction la plus adéquate pour les instances de test des mots ambigus. L'entrée de la méthode est constituée des résultats de la méthode de WSD pour ces instances, qui concernent des clusters de sens. La sortie de la méthode correspond à un EQV de traduction.

Dans les cas où le sens proposé pour une instance par la méthode de WSD est illustré par un cluster à **un seul EQV**, l'EQV en question correspond à la proposition de la méthode de sélection lexicale. La méthode de sélection lexicale intervient uniquement dans les cas où la proposition de la méthode de WSD consiste en un cluster à **plusieurs EQVs**. Elle sert ainsi à **filtrer** le cluster afin de sélectionner l'EQV qui constituerait la traduction la plus adéquate de la nouvelle instance du mot ambigu dans le contexte de la LC ; autrement dit, l'EQV qui serait le mieux « intégré » dans la traduction. Il faut souligner que l'ensemble des résultats de la WSD pour toutes les instances est traité par la méthode de sélection lexicale et non pas seulement les résultats qui correspondent aux instances correctement désambiguïsées.

3.2. Vers une évaluation pondérée des résultats de sélection lexicale

3.2.1. Principes de l'évaluation

Les propositions effectuées par la méthode de sélection lexicale peuvent être pondérées en tenant en compte des relations entre les EQVs clustérisés. La performance de la méthode est donc estimée sur la base des scores attribués. Cette **évaluation pondérée** est en accord avec la suggestion de Resnik et Yarowsky (1997, 2000) à propos de la pondération différenciée des erreurs de désambiguïsation, lors de l'évaluation de la performance des systèmes de WSD[12] : des propositions erronées pouvant provoquer un malentendu devraient

[12] Selon Resnik et Yarowsky, cette pénalisation différenciée pourrait se fonder sur la **distance fonctionnelle communicative**, qui serait définie dans un cadre psycholinguistique ou relativement

être davantage pénalisées que celles ne correspondant pas exactement à la réponse souhaitée, mais qui lui sont proches selon certains critères.

Nous estimons que cette pénalisation différenciée pourrait s'avérer bénéfique dans un cadre de sélection lexicale pour la TA. La sélection d'un mot différent de la référence devrait être pénalisée différemment selon le **degré de leur proximité sémantique**. Ainsi, la sélection d'un EQV sémantiquement similaire à l'EQV de référence devrait être moins pénalisée que celle d'un EQV véhiculant un sens différent. Cette possibilité d'une pondération différenciée des erreurs de traduction n'est pas prise en compte par les métriques existantes, en raison de leur incapacité à traiter des solutions sémantiquement pertinentes ne correspondant pas à la référence.

L'évaluation de notre méthode de sélection lexicale se fonde donc sur les principes suivants :

1. lorsque le cluster proposé par la méthode de WSD pour une instance contient la traduction de référence :

a. la proposition par la méthode de sélection lexicale de l'EQV correspondant à la référence est considérée comme **correcte** et obtient un score égal à 1

b. la proposition d'un autre EQV du cluster est considérée comme **correcte** et obtient un score proportionnel à la proximité sémantique de l'EQV sélectionné à la référence.

2. lorsque le cluster proposé ne contient pas la traduction de référence : une traduction est proposée pour l'instance en question mais la proposition est considérée comme **erronée**, dans la mesure où le cluster décrit un sens qui diffère de celui de la référence.

Le critère 1(a) est basé sur la notion de **précision**, dans le sens qu'il considère comme correcte une proposition qui correspond à la référence de manière exacte. En revanche, le critère 1(b) permet une « flexibilité » de l'évaluation, en considérant comme correctes des propositions sémantiquement

à une tâche précise. Ils soutiennent que, dans la TA, seules les distinctions sémantiques lexicalisées différemment dans la LC devraient être pénalisées. Ce principe permettrait de ne pas prendre en compte des distinctions sémantiques monolingues non pertinentes pour la traduction. Néanmoins, il ne permettrait pas de traiter la similarité sémantique des EQVs de traduction.

pertinentes ne correspondant néanmoins pas à la référence de manière exacte. Ce critère d'évaluation pourrait donc être considéré comme fondé sur une notion de **précision flexible** et **enrichie**.

3.2.2. Pondération des propositions de sélection lexicale

Dans le cas d'une correspondance exacte entre la traduction proposée par la méthode de sélection lexicale et la traduction de référence, la proposition est considérée comme **absolument correcte**. C'est le cas, par exemple, de la proposition faite pour l'instance de *movement* présente dans l'unité de traduction suivante :

> Of course, it is particularly important for people with special needs to have access to means of transport, services and all types of installation if they are to exercise their right to freedom of **movement.**
>
> Βεβαίως, είναι ιδιαίτερα σημαντικό να υπάρξει πρόσβαση των ατόμων με ειδικές ανάγκες στα μέσα μεταφοράς, στις υπηρεσίες, στις κάθε είδους εγκαταστάσεις, ώστε τα άτομα με ειδικές ανάγκες να έχουν τη δυνατότητα της ελεύθερης **κυκλοφορίας**.

La prédiction de la méthode de WSD pour cette instance consiste en le cluster suivant : {*κυκλοφορία, κίνηση, διακίνηση*}. La méthode de sélection lexicale filtre ce cluster, en s'appuyant sur le contexte de la LC, et propose l'EQV *κυκλοφορία* (kykloforia) comme traduction. Cette proposition correspond à la référence ; elle est considérée comme absolument correcte et obtient, par conséquent, un score égal à 1.

Dans le cas où un EQV différent de la référence est sélectionné, sa pertinence sémantique est proportionnelle au degré de sa proximité à l'EQV de référence. Le score attribué à une telle proposition ne correspondant pas à la référence de manière exacte mais entretenant une relation de similarité avec elle (étant trouvée dans le même cluster) est calculé sur la base du **score de similarité** entre l'EQV proposé et la traduction de référence dans les **contextes de la LC**. Ce score est fourni lors du calcul de similarité sémantique entre les EQVs (cf. §2.3.3, chapitre 6). Un EQV similaire à la traduction de référence est proposé pour l'instance de *movement* présente dans l'unité de traduction suivante :

Some speakers here have dwelt, quite understandably, on the problems that would arise from the fact that Iceland and Norway, through not being Member Stated of the European Union, will be excluded from discussions on the free **movement** area if it were to be incorporated into the Union treaties.

Ορισμένοι ομιλητές αναφέρθηκαν, και ορθά, στα προβλήματα που θα προκύψουν από το γεγονός ότι η Ισλανδία και η Νορβηγία, λόγω του ότι δεν είναι μέλη της Ευρωπαϊκής Ένωσης, θα αποκλειστούν από τις συζητήσεις για τη μελλοντική ανάπτυξη του τομέα της ελεύθερης **διακίνησης** εάν ο τομέας αυτός θα περιλαβανόταν στις Συνθήκες της Ένωσης.

Le sens proposé par la méthode de WSD pour cette instance de *movement* est décrit par le cluster {*κυκλοφορία, κίνηση, διακίνηση*} et est le sens de « mouvement physique ». Ce cluster est filtré lors de la sélection lexicale et la traduction proposée est *κυκλοφορία* (kykloforia), dont l'utilisation dans le texte de la traduction, à la place de *διακίνηση* (diakinisi), serait correcte. La proposition est donc considérée comme correcte, l'EQV se trouvant dans le même cluster que la référence, et la proposition obtient un score égal au score de similarité de l'EQV (*κυκλοφορία*) avec la traduction de référence (*διακίνηση*) dans les contextes de la LC : 0,098.

La pondération des propositions faites par la méthode de sélection lexicale pour les instances d'un mot ambigu *M* se fait par l'algorithme décrit dans la figure 6.

```
compteur_prédictions_correctes ← 0
clusters_M ← renvoie_liste_clusters_M
table_similarité_LC ← renvoie_table_similarité_LC_M

Pour chaque instance de M
        traduction_référence = trouve_traduction_référence(instance)
        prédiction = fait_prédiction(instance)

        Si prédiction = traduction_référence
                incrémente(compteur_prédictions_correctes)
        FinSi
        SiNon
                vrai = 0
                Pour chaque cluster de clusters_ M
                        Si {prédiction, traduction_référence} Є cluster
                                retourne vrai = 1
                                exit boucle
                        FinSi
                FinPour
                Si vrai = 1
                        score_similarité = table_similarité_LC[prédiction, traduction_référence]
                        compteur_prédictions_correctes +=score_similarité
                FinSi
        FinSiNon
FinPour
```

Figure 6. Algorithme de pondération des prédictions de traduction

L'algorithme prend en entrée la liste des clusters du mot ambigu *M* et sa table de similarité dans la LC. Pour chaque instance de test de *M*, la traduction de référence est trouvée et une prédiction de traduction est faite par la méthode de sélection lexicale. L'algorithme attribue un poids à la traduction sélectionnée, qui dépend de sa correspondance ou de sa proximité à la référence.

3.2.3. Avantages de l'évaluation pondérée

Les propositions effectuées par la méthode de WSD dépendent des informations de la LS retenues à partir du corpus d'apprentissage. En revanche, celles de la méthode de sélection lexicale, qui se fondent sur les résultats de la méthode de WSD, dépendent fortement des informations de la LC issues du corpus d'apprentissage.

Les informations contextuelles source et cible relatives à un EQV (traduction de référence) dans le corpus de test peuvent diverger de manière importante des informations retenues pour cet EQV à partir du corpus d'apprentissage. Néanmoins, cette divergence n'empêche pas notre méthode de proposer des traductions alternatives sémantiquement pertinentes, plus ou moins fortement liées à l'EQV de référence. Cette propriété rend le processus de sélection lexicale plus flexible et plus performant que lorsque des correspondances inter-langues sont établies entre des informations contextuelles source, relatives au mot ambigu, et un seul de ses EQVs de traduction. De telles correspondances sont retrouvées, par exemple, dans les travaux de désambiguïsation et de TA où des relations biunivoques sont modélisées entre sens et EQV, c'est-à-dire où un EQV est supposé correspondre à un sens distinct du mot ambigu.

3.2.4. Comparaison avec d'autres métriques d'évaluation

La seule des métriques d'évaluation de la TA existantes qui prend en compte les relations de similarité sémantique entre les mots d'une traduction automatiquement générée et ceux trouvés dans la traduction de référence est METEOR (Banerjee et Lavie, 2005 ; Lavie et Agarwal, 2007). Cette métrique est fondée sur un concept généralisé de correspondance d'unigrammes entre les traductions : des unigrammes peuvent être mis en correspondance sur la base de leurs **formes de surface**, de leurs **formes tronquées** et de leurs **sens**.

METEOR évalue une traduction en calculant un score fondé sur les correspondances entre les mots de cette traduction et les mots de la traduction de référence[13]. Lors de la comparaison de deux chaînes de caractères, METEOR crée un alignement de mots en mettant en correspondance un mot de chaque chaîne avec un mot de l'autre. Cet alignement est produit de manière incrémentale par une séquence de modules : le module « exact », qui met en correspondance deux mots s'ils sont exactement les mêmes ; le module « porter stem », qui fait correspondre deux mots variantes morphologiques l'un de l'autre (c'est-à-dire

[13] Lorsque plusieurs traductions de référence sont disponibles, la traduction proposée par le système obtient un score indépendamment par rapport à chacune d'elles, et la paire de traductions qui obtient le meilleur score est retenue.

ayant la même racine après avoir été tronqués) ; le module de « synonymie WN » (WN synonymy), qui met en correspondance deux mots considérés comme synonymes, c'est-à-dire appartenant au même synset de WordNet. Ce dernier module permet de prendre en compte les relations de similarité sémantique entre les mots de la traduction proposée et les mots de la traduction de référence. Il est montré que l'exploitation de ces relations augmente la corrélation entre les résultats de la métrique et des jugements humains sur la qualité de la traduction (Banerjee et Lavie, *ibid*.).

Cependant, l'utilisation de METEOR pour d'autres langues que l'anglais nécessite l'adaptation de certains de ses aspects qui sont dépendants des langues, dont un concerne les modules de correspondance de mots. L'utilisation de WordNet rend en effet impossible l'utilisation de la métrique en question pour des langues où un tel inventaire n'est pas disponible. Ce problème est d'ailleurs rapporté par Lavie et Agarwal (*ibid*.), qui soulignent que les versions de METEOR pour l'espagnol, le français et l'allemand contiennent seulement les modules de correspondance exacte et de troncation, en raison de la non disponibilité publique de ressources de type WordNet pour ces langues. Afin de se passer de la nécessité de telles ressources, ils envisagent la possibilité de développer de nouveaux modules de synonymie pour d'autres langues en se basant sur des méthodes alternatives.

Une autre faiblesse de cette métrique concerne le fait que le module de synonymie met en correspondance deux mots si au moins un de leurs sens appartient au même synset de WordNet. Ce principe implique qu'au moins un sens de chaque mot représente le même concept qu'un sens de l'autre. Néanmoins, cet algorithme de détection de la synonymie est trop simpliste, dans la mesure où il ne désambiguïse pas les mots avant de tester s'il s'agit de synonymes.

Notre métrique d'évaluation se fonde sur la même idée que METEOR, à savoir la prise en compte des **correspondances entre mots sémantiquement similaires**. Néanmoins, elle ne nécessite pas de ressources sémantiques prédéfinies (comme WordNet). Les relations sémantiques entre les EQVs sont extraites automatiquement du corpus d'entraînement par la méthode d'acquisition de sens, qui est indépendante de langue. Par conséquent, notre

métrique serait utilisable pour d'autres langues. La seule condition serait que la méthode d'acquisition de sens soit d'abord appliquée sur un corpus d'entraînement parallèle des langues en question. Par exemple, si la traduction est faite du français vers l'espagnol, il faudrait un corpus d'entraînement français-espagnol afin d'acquérir les clusters des EQVs de traduction espagnols des mots français. En revanche, si la direction de traduction était inversée (espagnol-français), l'acquisition des clusters des EQVs français des mots espagnols serait possible en appliquant la méthode d'acquisition de sens dans cette direction.

3.3. Métrique de précision enrichie *vs* métrique de précision stricte

Les résultats obtenus en utilisant la métrique d'évaluation enrichie, que nous venons de présenter, ont été comparés à ceux obtenus en utilisant une **métrique de précision stricte**.

Les scores de l'évaluation pondérée sont calculés selon les critères présentés dans le paragraphe 3.2.1. En revanche, la métrique de précision stricte calcule les scores en s'appuyant sur le **critère strict de précision** (principe 1(a), cf. §3.2.1.). Selon ce critère, les propositions de traduction sont considérées comme correctes **si et seulement si** elles correspondent à la traduction de référence de manière exacte.

3.4. Métrique de précision enrichie *vs* méthode de base

Nous avons également comparé les résultats obtenus lors de l'évaluation enrichie à ceux d'une **méthode de base** (*baseline*), qui consiste à sélectionner l'**EQV le plus fréquent** d'un mot ambigu comme traduction de toutes ses nouvelles instances. Cette heuristique s'avère en effet très puissante dans le cadre de cette évaluation : lorsque l'EQV le plus fréquent dans le corpus d'entraînement est également le plus fréquent dans le corpus de test (ce qui est le cas pour la quasi-totalité des mots ambigus étudiés), le nombre de propositions correctes effectuées par la méthode de base est élevé et elles se voient toutes

attribué un score de 1. En revanche, lorsque notre méthode propose des traductions sémantiquement similaires à la référence, ces prédictions sont pondérées sur la base du score de similarité entre la traduction proposée et la référence. Ce score est donc toujours inférieur à 1, les deux mots n'étant jamais des synonymes absolus. C'est la raison pour laquelle l'amélioration apportée par notre méthode n'est pas proportionnelle à la différence observée au niveau des résultats quantitatifs.

Il faut également noter que, contrairement à la méthode de sélection, la méthode de base n'est pas vulnérable au bruit présent dans les données de la LC et qui provient du corpus de test. Ce bruit consiste en des erreurs d'étiquetage morphosyntaxique et de lemmatisation (comme nous l'avons déjà expliqué dans le §2.2.3. du chapitre 4). Enfin, il est également important de souligner que la méthode de sélection se base sur les résultats de la méthode de WSD ; et que, les erreurs survenues lors de cette étape se répercutent par conséquent sur la sélection lexicale.

Dans le paragraphe suivant, nous présenterons les résultats de notre méthode calculés à l'aide des deux métriques utilisées, ainsi que ceux obtenus par la méthode de base.

3.5. Evaluation quantitative

3.5.1. Mesures utilisées

Comme dans le cas de l'évaluation de la méthode de WSD, la performance de la méthode de sélection lexicale est mesurée à l'aide des mesures de **rappel**, de **précision** et de **f-mesure**. Ces scores sont, d'abord, calculés selon le critère de précision stricte puis, selon le critère de précision enrichie. Dans le premier cas, nous parlons donc d'une **évaluation stricte** et, dans le second, d'une **évaluation enrichie**.

A noter que le **rappel** est calculé **deux fois** dans le cadre de chacune de ces évaluations :

- une première fois, relativement au **nombre de prédictions** fournies par la méthode de **WSD** (exploitées par la méthode de sélection pour proposer des traductions correctes)
- une autre fois, relativement au **nombre total** des nouvelles **instances** du mot ambigu.

Ainsi, le premier score de rappel (‘**rappel1’**) correspond au rapport du nombre de sélections de traduction correctes au nombre des prédictions de désambiguïsation et le deuxième (‘**rappel2’**), correspond au rapport du nombre de sélections correctes au nombre des nouvelles instances du mot ambigu. Dans le deuxième cas le rappel diminue, dans la mesure où des propositions de traductions ne sont pas faites pour les instances non désambiguïsées.

La **précision** correspond, quant à elle, au rapport du nombre de sélections correctes au nombre de sélections lexicales faites par le système.

La **f-mesure** combine les valeurs de rappel et de précision en une valeur unique (cf. §2.3). Deux valeurs de la f-mesure sont ainsi calculées (‘**f-mesure1**’, ‘**f-mesure2**’) lors de chaque évaluation, correspondant aux deux scores de rappel (‘**rappel1’**, ‘**rappel2’**).

En outre, nous calculons le score de la méthode de base qui, comme dans le cas de la WSD, réunit les scores de rappel et de précision. Deux valeurs sont également calculées pour cette méthode : une relativement au nombre des prédictions de WSD (‘**baseline1’**) et une autre relativement au nombre total des instances de test (‘**baseline2’**).

3.5.2. Résultats pour les mots du lexique manuellement généré

3.5.2.1. Comparaison des scores obtenus

Dans ce paragraphe, nous présentons les résultats de la méthode de sélection lexicale obtenus dans le cadre des évaluations stricte et enrichie, et ce, pour la totalité des mots ambigus du **lexique construit à la main**. Ces résultats sont comparés à ceux fournis par la méthode de base. Dans la figure 7, nous

présentons les scores des 'f-mesures' 2 et de la méthode de base, calculés séparément pour chaque mot du lexique manuellement généré.

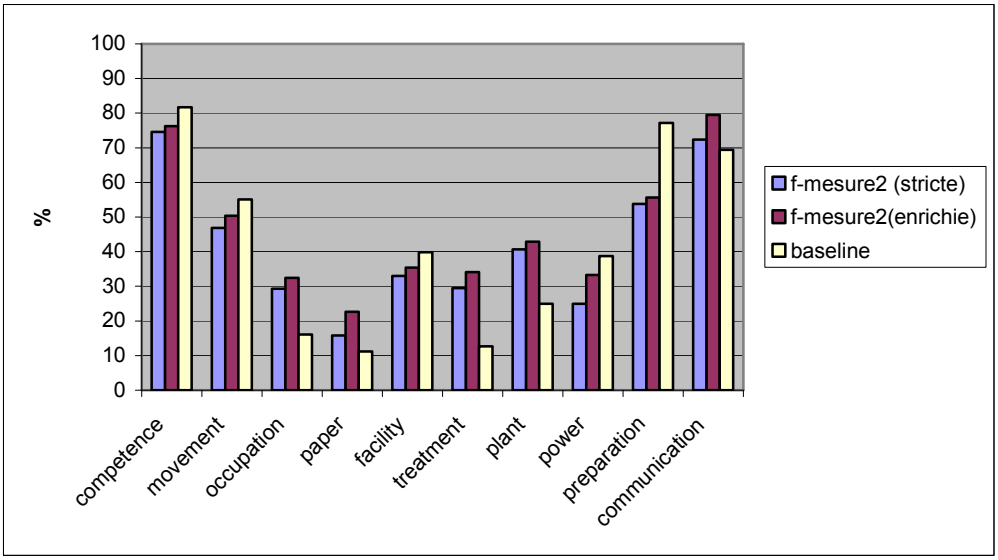


Figure 7. Evaluation de la sélection lexicale pour chaque mot du lexique manuellement généré

Des divergences assez remarquables sont observées entre mots différents, aussi bien en ce qui concerne la performance des deux méthodes (c'est-à-dire, le taux des instances pour lesquelles une prédiction est faite), que les scores obtenus. Afin d'avoir une vue d'ensemble de la performance de notre méthode, mesurée dans le cadre des deux évaluations, nous présentons les résultats globaux, obtenus pour la totalité des mots étudiés. Les 'f-mesures' 1 et 2 de la méthode de sélection lexicale, obtenues lors de l'évaluation stricte, sont comparées aux 'f-mesures' 1 et 2, obtenues lors de l'évaluation enrichie, au sein de la figure 8. Les scores obtenus par ces métriques sont également comparés à ceux de la méthode de base.

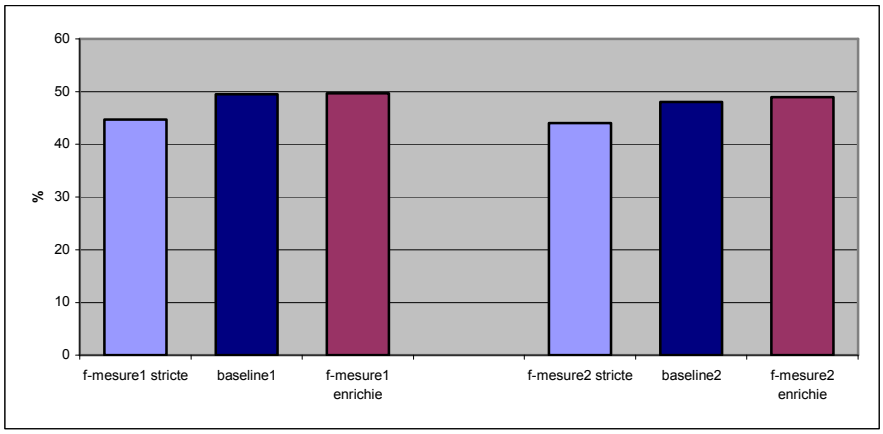


Figure 8. Evaluation de la sélection lexicale pour la totalité du lexique manuellement généré

La 'f-mesure 1 stricte' est de 44,73% tandis que la 'f-mesure 1 enrichie' est de 49,65%. En outre, la 'f-mesure 2 stricte' est de 44,08% et la 'f-mesure 2 enrichie' est de 48,92%. La divergence des scores entre 'f-mesure 1' et 'f-mesure 2', dans le cadre de chaque évaluation, s'explique par le nombre plus élevé des instances de test considérées lors du calcul du deuxième score de rappel (14.456 vs. 14.036). Cette divergence est due également au fait que le nombre total d'instances de test du mot ambigu, par rapport auquel le 'rappel2' est calculé, comprend des instances non désambiguïsées, qui ne sont donc pas traitées par la méthode de sélection lexicale. En revanche, le 'rappel1' et, par conséquent, la 'f-mesure1', sont calculés sur la base du nombre des prédictions de WSD ; 14.036 instances de test ont été désambiguïsées et une sélection lexicale a été effectuée pour la totalité de ces instances.

Cette divergence s'observe également dans les scores de la méthode de base ('baseline1' : 49,5% et 'baseline2' : 48,06%), mais est due ici uniquement à la différence du nombre des instances de test considérées lors du calcul des scores. Ce nombre est plus élevé lors du calcul du second score et, tel que cela apparaît dans le dénominateur de la formule qui sert à calculer le score, le score fourni est plus faible.

3.5.2.2. Evaluation stricte vs. evaluation enrichie

Nous observons donc que les scores attribués à la méthode de sélection lexicale lors de l'évaluation enrichie dépassent de manière assez nette ceux obtenus lors de l'évaluation stricte mais représentent une amélioration très faible par rapport aux résultats de la méthode de base. Ces scores sont-ils représentatifs de la performance réelle des méthodes utilisées ? Nous avons remarqué que tel est le cas lors de la comparaison entre les résultats des mesures strictes et les scores baseline, mais pas lors de la comparaison entre ces derniers et les résultats des mesures enrichies. Cette évaluation devrait pourtant permettre de prendre en compte les améliorations apportées par cette méthode.

Au premier regard porté sur les résultats de l'évaluation, il semble que la différence entre les résultats des mesures stricte et enrichie indique le taux des propositions de traduction sémantiquement pertinentes, différentes de la référence. Néanmoins, nous avons observé que cette différence ne reflète pas réellement la divergence entre le nombre de propositions correspondant à la référence et celui qui comprend, en outre, les propositions proches de la référence.

Effectivement, ces deux nombres diffèrent largement : dans le premier cas, 6.279 propositions correctes sont faites tandis que, dans le second, les propositions correctes s'élèvent à 10.473. Cette divergence montre que 4.194 traductions similaires à la référence sont proposées par la méthode de sélection, sur un nombre total de 14.456 d'instances de test. Cette différence correspond à une augmentation de 29,01% du nombre d'instances pour lesquelles une traduction pertinente est proposée. Pourquoi une telle différence n'est-elle pas reflétée au niveau des scores ?

La raison en est la manière dont les propositions similaires à la référence sont pondérées lors de l'évaluation enrichie. Ces propositions sont en effet **toujours** pondérées par un **score inférieur à 1**. Etant donné que le rappel et la précision sont calculés sur la base des scores attribués aux propositions et non sur la base du nombre de propositions faites, la différence entre les scores de la f-mesure, lors des deux évaluations, n'est pas proportionnelle à la quantité de propositions pertinentes. Elle permet, néanmoins, d'avoir une idée de

l'amélioration apportée par la méthode de sélection, lors de la prise en compte des propositions similaires à la référence.

3.5.2.3. Evaluations stricte et enrichie vs. évaluation de la méthode de base

Analysons maintenant les scores fournis par les évaluations stricte et enrichie par rapport à ceux de la méthode de base. Le fait que le score de la mesure stricte soit inférieur à la baseline, montre que pour une partie des instances dont la traduction dans le corpus de test correspond à l'EQV le plus fréquent (cas pris en compte par la baseline), la méthode de sélection ne propose pas cet EQV précis. Dans ces cas, deux possibilités existent : soit aucune proposition n'est faite, soit des traductions sémantiquement similaires à la référence sont proposées. Dans ce dernier cas, les propositions sont prises en compte par la métrique enrichie. Cependant, leurs scores sont inférieurs à ceux attribués aux propositions de la baseline. Regardons ce qu'indique la comparaison entre les scores enrichis et les scores de la baseline.

La différence entre ces scores semble très petite dans le cadre de la première évaluation et légèrement plus importante, dans le cas de la deuxième. Est-ce que les divergences observées sont représentatives de la quantité des propositions pertinentes supplémentaires effectuées par la méthode de sélection ? Suite à une observation plus attentive des résultats de la méthode, nous pouvons répondre par la négative. En effet, cette méthode propose, au total, 10.473 traductions pertinentes (les prédictions sémantiquement similaires à la référence incluses), tandis que la méthode de base en propose uniquement 6.948 ; d'où une différence de 3.525 sur un nombre total d'instances de test de 14.456, et qui correspond à une augmentation de 24,38% du nombre d'instances pour lesquelles une traduction pertinente est proposée.

La différence quant au nombre de traductions proposées est donc très importante et elle n'est aucunement reflétée dans la différence entre les scores attribués. La raison en est, ici aussi, que les propositions sémantiquement liées à la référence obtiennent toujours un score inférieur à 1, contrairement aux propositions de la méthode de base qui sont toujours pondérées par 1.

3.5.3. Résultats pour les mots du lexique automatiquement généré

Dans ce paragraphe, nous présentons les résultats de la méthode de sélection lexicale obtenus dans le cadre des évaluations stricte et enrichie, pour la totalité des mots trouvés dans le **lexique automatiquement généré**. Les résultats sont comparés à ceux de la méthode de base sur les mêmes données. Ici aussi, les 'f-mesures' 1 et 2, obtenues lors des deux étapes de l'évaluation (stricte et enrichie), sont comparées aux scores 'baseline1' et 'baseline2'. Les prédictions de WSD concernent 174.841 instances de test et des prédictions de sélection lexicale sont faites pour la totalité de ces instances. Les résultats des évaluations stricte et enrichie sont comparés à ceux de la méthode de base au sein de la figure 9.

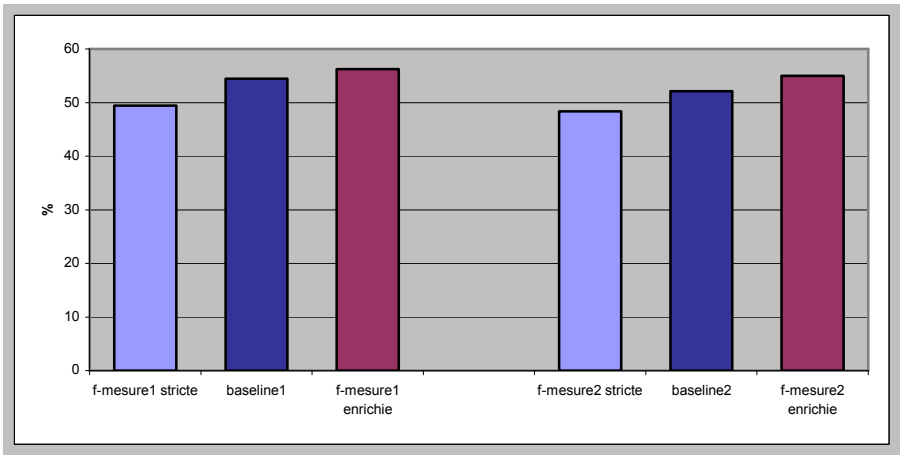


Figure 9. Evaluation de la sélection lexicale pour le lexique automatiquement construit

Comme dans le cas du lexique manuellement généré, les résultats de la méthode de sélection lors de l'évaluation enrichie ('f-mesure1 enrichie' : 56,22%, 'f-mesure2 enrichie' : 55,01%) dépassent les résultats calculés lors de l'évaluation stricte ('f-mesure1 stricte' : 49,43%, 'f-mesure2 stricte' : 48,37%). Ici aussi, ces derniers sont inférieurs aux scores de la méthode de base ('baseline1' : 54,42%, 'baseline2' : 52,14%).

Malgré le biais introduit dans la comparaison quantitative en faveur des scores de la méthode de base, nous observons une amélioration apportée par notre méthode, dans le cas de ce lexique. Néanmoins, comme dans le cas du

lexique manuellement généré, la différence observée au niveau de la quantité des propositions faites est beaucoup plus importante que celle observée au niveau des scores. Ainsi, notre méthode propose 138.994 traductions sémantiquement pertinentes, tandis que la méthode de base en propose 95.154 ; d'où une différence de 43.840 sur un nombre total d'instances de test de 182.485, et qui correspond à une augmentation de 24,02% du nombre d'instances pour lesquelles une traduction pertinente est proposée.

Même si cela n'est donc pas évident en considérant les résultats de l'évaluation, notre méthode de sélection lexicale permet une augmentation très importante du nombre de traductions proposées. Cette augmentation quantitative se conjugue par ailleurs au bénéfice qualitatif que représente le fait de proposer des traductions alternatives pertinentes – d'où une augmentation de la richesse du vocabulaire utilisé dans la traduction. Envisager des manières permettant à ces différences d'apparaître de façon plus évidente au niveau des résultats quantitatifs de l'évaluation fait partie des perspectives de ce travail.

CONCLUSION

L'évaluation de la méthode de WSD, qui exploite les résultats de notre méthode d'acquisition de sens, démontre que la prise en compte des clusters d'EQVs améliore de manière significative la performance de la désambiguïsation par rapport à la baseline. La méthode de base utilisée ici consiste à sélectionner l'EQV le plus fréquent d'un mot ambigu comme illustrant le sens de toutes ses nouvelles instances, à l'instar de ce qui est fait dans les tâches d'évaluation multilingues des systèmes de WSD. La précision et le rappel de notre méthode sont très élevés, ainsi que sa couverture.

La méthode de sélection lexicale, qui opère sur les résultats de la désambiguïsation, rend possible la proposition de traductions sémantiquement proches de la référence. Cette méthode augmente considérablement la quantité de traductions proposées, amélioration qui n'est pourtant pas reflétée au niveau des scores obtenus.

La métrique d'évaluation enrichie proposée, même si elle ne reflète pas l'étendue de cette amélioration, permet néanmoins d'indiquer qu'une amélioration est survenue, surtout par rapport aux résultats de la méthode stricte. La métrique enrichie prend en compte les relations de similarité sémantique existant entre les équivalents de traduction des mots. Contrairement à d'autres métriques proposées, qui tiennent compte de ces relations en s'appuyant sur des inventaires prédéfinis, la métrique d'évaluation proposée ici ne nécessite pas le recours à de telles ressources, mais exploite les résultats de la méthode d'acquisition de sens. Les informations sémantiques étant obtenues de manière automatique à partir de corpus, la métrique s'avère donc utilisable pour toute évaluation impliquant d'autres langues.

CONCLUSION

1. Bilan

La recherche menée dans le cadre de ce travail a traité la question de l'acquisition de sens pour la désambiguïsation lexicale dans un cadre de Traduction. Une grande partie de l'analyse au niveau théorique a été consacrée à l'étude de l'applicabilité **des hypothèses distributionnelles** du **sens** et de la **similarité sémantique** dans un cadre bilingue. Ces hypothèses étant sous-jacentes à maintes méthodes d'analyse sémantique s'appuyant sur des informations monolingues, nous montrons qu'elles peuvent être également exploitées dans un cadre impliquant des langues différentes, permettant aux méthodes qui s'en servent de fournir des résultats pertinents.

Les hypothèses distributionnelles en question constituent ainsi les principes sur lesquels s'appuient les méthodes implémentées. Dans notre méthode d'acquisition de sens, elles ont été couplées avec l'hypothèse d'une **correspondance sémantique** entre mots en relation de traduction. L'analyse effectuée par cette méthode met au jour des relations de similarité sémantique entre les équivalents des mots ambigus de la langue source. Ces relations servent au clustering sémantique des équivalents ; les clusters générés de cette manière sont projetés sur les mots ambigus source, projection qui rend possible l'induction de distinctions sémantiques au sein de ces mots.

Les sens mis en évidence par cette méthode, dans la mesure où ils sont représentés par des correspondances sémantiques inter-langues, s'avèrent pertinents pour la désambiguïsation dans un cadre de traduction. Contrairement aux méthodes établissant des correspondances biunivoques entre sens et équivalents, les correspondances sont ici élaborées sur la base des relations paradigmatiques repérées entre les équivalents, ce qui constitue une des originalités de cette représentation sémantique. Cette particularité de la méthode lui permet, à la fois, de capter de véritables correspondances de sens et de prendre en compte une caractéristique essentielle à tout processus de traduction, présente dans les textes traduits : celle qui consiste à recourir, de la part du traducteur, à des mots sémantiquement similaires pour traduire le sens d'un mot ambigu source, dans un contexte donné.

Notre méthode étant non supervisée et s'appuyant sur des informations distributionnelles et de traduction extraites de corpus, est indépendante de la langue. Ne nécessitant pas de données pré-étiquetées, cette méthode peut facilement être appliquée à des paires de langues différentes, à la condition que des corpus parallèles soient disponibles pour l'entraînement et qu'existent des outils permettant leur lemmatisation et leur étiquetage morphosyntaxique. En outre, comme elle est totalement dirigée par les données, notre méthode permet l'élaboration automatique d'inventaires de sens relatifs aux données traitées.

L'inventaire ainsi généré se différencie de manière importante des inventaires « classiques », d'une part, par sa structure et d'autre part, par son contenu. Au sein de cet inventaire, des **liens** entre **sens lexicaux** sont modélisés, liens pouvant être exploités pour la **modification de la granularité** des descriptions sémantiques fournies. De cette manière, en utilisant uniquement des informations internes, la méthode parvient à révéler des sens lexicaux plus ou moins grossiers. La modélisation sémantique effectuée est donc **dynamique**, dans la mesure où elle permet d'accéder à des informations de quantité et de qualité variables, en fonction des besoins qui se présentent dans un cadre donné.

Par ailleurs, étant donné leur caractère inter-langue, les descriptions incluses au sein de l'inventaire de sens peuvent être exploitées pour la désambiguïsation et la sélection lexicale dans un cadre de traduction. Ces deux méthodes, de désambiguïsation et de sélection lexicale, ont été mises en place dans le cadre de

ce travail et exploitent le contenu de l'inventaire que nous avons élaboré. En ce qui concerne le fonctionnement des méthodes, les tâches de désambiguïsation et de sélection lexicale sont assimilées lorsque le sens sélectionné pour un mot est représenté par un cluster contenant un seul équivalent, et distinctes, lorsque ce sens est représenté par un cluster plus étendu.

Les avantages liés à la nature de la modélisation de sens que nous proposons se sont également manifestes au niveau de l'évaluation des méthodes de désambiguïsation et de sélection lexicale. En effet, dans la mesure où elle met en évidence les relations sémantiques entre les équivalents des mots ambigus et où elle opère une distinction entre sens proches et distants, la modélisation proposée rend possible l'élaboration d'une **métrique d'évaluation pondérée**. Cette métrique se différencie des métriques classiques d'évaluation du résultat de la Traduction Automatique, ainsi que de celles utilisées pour évaluer le résultat des méthodes de désambiguïsation multilingue. Basée sur le principe de **précision enrichie**, notre métrique tient compte de la pertinence sémantique des résultats du processus de sélection lexicale.

Les conclusions sur l'évaluation quantitative des méthodes de désambiguïsation et de sélection lexicale, présentée dans le dernier chapitre de cette thèse, montrent en effet que la modélisation sémantique fournie par notre méthode d'acquisition de sens est bénéfique aux tâches de désambiguïsation et de sélection lexicale, opérant dans un cadre bilingue. Ainsi, les apports qualitatifs – qui consistent en une modélisation dynamique de la sémantique lexicale, permettant la prise en compte de relations inter-sens –, se conjuguent à une augmentation de la qualité de l'évaluation, qui permet, elle, la prise en compte de propositions sémantiquement similaires. Cette évaluation révèle l'amélioration des résultats apportée lors d'une tâche de sélection lexicale, par la méthode proposée, amélioration qui ne se voit pourtant pas clairement lors de la comparaison des résultats de la métrique d'évaluation enrichie à ceux fournis par une autre méthode de sélection et par une autre métrique d'évaluation plus simples. Néanmoins, cette amélioration pourrait devenir plus évidente par une modification du processus de pondération, sujet qui sera abordé au sein des perspectives.

Après avoir dressé le bilan des apports de cette étude aux niveaux théorique et pratique, nous allons désormais présenter les futures pistes de recherche envisagées.

2. Perspectives

2.1. D'un point de vue opératoire

2.1.1. Cooccurrences d'ordre plus élevé

La méthode d'acquisition de sens proposée exploite des **contextes lexicaux élargis**, ce qui permet de restreindre l'effet de la dispersion des données. Néanmoins, dans la mesure où elle s'appuie sur un calcul distributionnel qui utilise des informations de cooccurrence de premier ordre, elle demeure relativement vulnérable à l'impact de ce phénomène. Une des pistes de recherche futures consisterait à utiliser des informations de **cooccurrence d'un ordre plus élevé**. Les informations de ce type étant situées à un plus haut niveau d'abstraction, elles permettraient d'établir des relations entre contextes sémantiquement similaires mais composés de mots différents. Nous estimons que ce type d'informations aurait un effet bénéfique sur les résultats de notre méthode, à savoir, le fait de pouvoir regrouper des sens de granularité très fine.

2.2.2. Modification dans la manière d'appréhender le contexte

Nous souhaiterions, en outre, mener des expériences en modifiant la manière dont le contexte local est pris en compte par les méthodes proposées. Dans cette étude, le contexte est appréhendé en tant que « sac de mots », sans considération quant à l'ordre des mots ni à leurs relations au sein des phrases. Une des pistes de recherche futures consisterait donc à transformer nos méthodes de façon à ce qu'elles puissent prendre en compte la position des cooccurrents, d'une part, par rapport au mot ambigu au sein de la LS et, d'autre part, par rapport aux EQVs, au sein de la LC. Le premier cas concerne les méthodes

d'acquisition de sens et de désambiguïsation, tandis que le second aurait un impact sur la méthode de sélection lexicale. Ce type de modification permettrait de traiter, de façon diversifiée, les différentes instances des mots ambigus apparaissant dans la même phrase ; ce qui n'est pas possible avec une conception des cooccurrents comme sac de mots.

Nous aimerions, par ailleurs, étudier l'influence que des informations sur les relations syntaxiques des mots, au sein des phrases, pourraient produire sur les résultats. Dans ce travail, nous avons fait le choix de ne pas tenir compte de ces informations, ce qui rend nos méthodes applicables à des langues ne disposant pas d'outils nécessaires pour un tel traitement. Néanmoins, cette piste pourrait, à la fois, représenter d'éventuels bénéfices, lorsque de tels outils sont disponibles, et fournir du contenu intéressant pour une étude théorique de l'influence des informations syntaxiques des mots sur le processus d'acquisition de sens.

2.2.3. Définition du seuil de pertinence des relations sémantiques

Une autre question que nous souhaiterions explorer davantage au niveau opératoire concerne la définition d'un **seuil** plus pertinent pour l'estimation de la signifiance des relations sémantiques proposées. Nous avons souligné les limitations imposées par le seuil proposé, qui coïncide avec la moyenne des scores de similarité attribués aux paires d'équivalents d'un mot (cf. § 3.2.1, chapitre 6). Ces limitations sont d'autant plus évidentes que le nombre d'équivalents est petit, ce qui est essentiellement le cas lors de l'exploitation d'un lexique automatiquement construit. Plus précisément, la moyenne ne permet pas d'obtenir des résultats pertinents lorsque le nombre d'équivalents est égal à deux ou, même, à trois. L'utilisation de la moyenne résulte à un traitement uniforme de tous les mots ambigus, sans considération de leurs particularités à propos du nombre d'équivalents de traduction possibles. La prise en compte de ces particularités aurait un impact sur la quantité et la qualité des sens proposés. Nous estimons que ces particularités pourraient être prises en compte par la définition d'un **seuil local**, pour chaque mot traité.

2.2.4. Inversion de la direction de traduction

La méthode proposée étant **directionnelle**, il serait également intéressant de procéder à une analyse de la sémantique des mots de la langue cible, en inversant la direction de traduction. Cette étape permettrait d'obtenir une image plus complète de la sémantique des mots de cette langue et la création d'un réseau complexe de correspondances liant les unités lexicales à un niveau sémantique. Le résultat d'une telle analyse consisterait donc en une description plus exhaustive des relations existant entre les régions sémantiques des langues impliquées.

2.2.5 Amélioration de la qualité de l'alignement lexical

Nous avons, par ailleurs, souligné la forte dépendance de notre méthode d'acquisition de sens à la qualité de l'alignement lexical. Une amélioration de cet alignement produirait un impact positif sur la qualité du lexique bilingue automatiquement élaboré pour l'ensemble des mots du corpus. Et, l'amélioration de la qualité de ce lexique produirait, à son tour, un impact positif sur la qualité des descriptions sémantiques obtenues, qui seraient plus complètes (comme c'est le cas pour les descriptions obtenues à partir du lexique élaboré à la main). Cette amélioration rendrait également possible l'analyse d'un plus grand nombre de mots ambigus, ce qui constitue une condition importante à l'**intégration** de la **méthode de désambiguïsation** proposée dans un **système de Traduction Automatique Statistique** réel.

2.2.6. Evaluation humaine des résultats des méthodes proposées

Enfin, nous estimons d'une étape supplémentaire d'évaluation pourrait être effectuée pour chacune des méthodes proposées, qui impliquerait des évaluateurs humains. Les acteurs de ces évaluations devraient être bilingues, dans les deux langues concernées, ou être, du moins, traducteurs professionnels. Dans le cas de la méthode d'acquisition de sens, l'intérêt serait d'évaluer la pertinence des représentations sémantiques engendrées.

Lors d'une évaluation de ce type dans le cas de la désambiguïsation, l'objectif serait d'estimer la pertinence des sens proposés pour de nouvelles instances des mots ambigus en contexte. En revanche, dans le cas de la sélection lexicale, il s'agirait plutôt d'évaluer la possibilité de permutation entre la traduction proposée et la traduction de référence au sein du texte de la traduction.

2.2. D'un point de vue applicatif

2.2.1. Désambiguïsation lexicale pour la Traduction

L'intégration de notre méthode de désambiguïsation dans un système de Traduction Automatique Statistique constitue l'une des pistes les plus intéressantes pour l'exploitation du travail présenté au sein de cette thèse. Notre méthode de sélection lexicale constitue une approximation du processus de sélection effectué dans un système de traduction, et traite un **problème simplifié de traduction**. Cette intégration permettrait donc d'estimer le véritable apport de la modélisation sémantique et des méthodes de désambiguïsation et de sélection lexicale proposées dans un cadre de Traduction Automatique. Nous pensons qu'une telle intégration mettrait également en évidence des améliorations éventuelles des méthodes proposées. Précisons, néanmoins, que pour que cette intégration soit possible, un lexique de grande envergure, caractérisé par une précision et un rappel élevés, doit être disponible.

En outre, l'intégration de notre méthode de désambiguïsation dans un système de SMT permettrait d'estimer l'utilité de notre **métrique d'évaluation** pour une tâche plus complète. Il serait également possible de comparer les résultats de l'évaluation fournis par cette métrique aux résultats fournis par d'autres métriques, couramment utilisées pour l'évaluation de la performance des systèmes de SMT.

2.2.2. Métrique d'évaluation pondérée

A propos de notre métrique d'évaluation, nous estimons qu'elle pourrait être modifiée par des manières alternatives de pondérer les propositions sémantiquement pertinentes et ce, afin que la comparaison des résultats de l'évaluation soit plus représentative des améliorations réelles apportées par la méthode de sélection lexicale.

En effet, les améliorations quantitatives et qualitatives apportées par les méthodes de désambiguïsation et de sélection lexicale ne sont pas proportionnelles à l'amélioration observée dans les résultats. Comme nous l'avons expliqué, la raison en est la manière dont la pondération des propositions sémantiquement pertinentes est effectuée. Pour être plus sophistiquée, cette pondération pourrait notamment faire appel à un coefficient capable de prendre en compte non seulement le score de similarité de la proposition à la traduction de référence, mais aussi le nombre d'équivalents candidats du mot source, le nombre des sens du mot et leur distinctivité, ainsi que le nombre des équivalents similaires à la référence et les poids de ces relations. Un paramètre supplémentaire à considérer concerne l'augmentation importante du nombre de propositions de traduction faites par la méthode de sélection lexicale relativement à la baseline.

2.2.3. Application à d'autres paires de langues

Une autre piste de recherche possible consisterait à appliquer les méthodes proposées à d'autres paires de langues. Etant donné le caractère statistique des méthodes et la nature endogène des informations exploitées, une telle extension serait tout à fait envisageable. La mise en œuvre des expériences sur une autre paire de langues (par ex. anglais-français) pourrait probablement permettre l'extraction de lexiques bilingues plus satisfaisants, dans la mesure où l'alignement des mots pourrait profiter d'indices non présents entre l'anglais et le grec (comme les *cognats*, qui sont des mots apparentés de deux langues).

2.2.4. Création automatique de corpus sémantiquement étiquetés

Tout au long de ce travail, nous avons insisté sur la possibilité d'exploitation de notre inventaire de sens pour la désambiguïsation dans la traduction et nous avons démontré son utilité par les résultats obtenus avec notre méthode non-supervisée de WSD. Nous considérons que l'analyse d'un plus grand nombre de mots ambigus par la méthode d'acquisition de sens et la désambiguïsation à l'aide des résultats de cette analyse permettraient l'élaboration automatique de **corpus sémantiquement étiquetés**. Ces corpus pourraient ensuite constituer une ressource pour l'entraînement d'algorithmes supervisés de désambiguïsation dans un cadre bilingue. Les sens des instances des mots ambigus au sein des corpus seraient illustrés par les clusters de leurs équivalents fournis au sein de notre inventaire.

Un tel processus d'étiquetage non supervisé présenterait l'avantage de l'annotation sémantique à l'aide des équivalents de traduction, à savoir la possibilité d'étiquetage d'un corpus d'une langue ne disposant pas d'inventaires sémantiques prédéfinis. Notre méthode présente même un certain avantage par rapport à certaines méthodes développées dans ce but (cf. §2.1.2.2, chapitre 2), puisqu'elle n'implique pas la nécessité de recourir à un inventaire de sens et ce, pour **aucune** des langues impliquées. En prenant en compte la **distance sémantique** entre les traductions d'un mot source à l'aide du processus de clustering –qui coïncide précisément à l'une des perspectives d'amélioration proposées pour les méthodes existantes – notre méthode permettrait l'étiquetage de mots source en fonction de leurs sens différents. Ne nécessitant pas de données annotées pour l'apprentissage, elle serait également adaptable à des corpus variés, dans la mesure où son lien au corpus d'entraînement est beaucoup plus faible que dans le cas d'une méthode supervisée. Nous considérons donc que l'investigation de cette piste présenterait un grand intérêt.

2.2.5. Exploitation de l'inventaire de sens pour la RIM

Enfin, il serait envisageable d'exploiter les représentations sémantiques obtenues dans un cadre de Recherche d'Information Multilingue (RIM). L'intérêt

de l'utilisation d'une ressource sémantique multilingue dans un tel cadre a déjà été souligné (cf. §2.1.2, chapitre 9). Néanmoins, les distinctions sémantiques fournies au sein des ressources prédéfinies sont parfois trop fines pour les besoins de la RIM, dans la mesure où elles ne renvoient pas toujours à des topics ou à des types de documents différents. C'est pourquoi des méthodes de clustering ont été proposées pour le regroupement des sens en question. Les clusters obtenus ne regroupent cependant pas toujours des mots sémantiquement similaires.

Les clusters fournis par notre méthode pour un mot ambigu se caractérisent précisément par la similarité sémantique des équivalents qu'ils contiennent. Nous estimons que, étant donné la possibilité de modification de la granularité des sens obtenus en fonction des besoins de l'application visée en matière de désambiguïsation, il serait intéressant d'envisager l'exploitation de nos sens de granularité grossière (conduisant à des topics différents) dans un cadre de RIM. La désambiguïsation des mots à l'aide des clusters qui leur sont associés présenterait les mêmes avantages que l'utilisation d'un réseau sémantique dans ce cadre, à savoir l'augmentation de la précision (en raison de la désambiguïsation) et du rappel (par référence aux ensembles de mots sémantiquement similaires dans la LC).

REFERENCES

AGIRRE, Eneko & SOROA, Aitor (2007a). *SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 7-12.

AGIRRE, Eneko & SOROA, Aitor (2007b). *UBC-AS: A Graph Based Unsupervised System for Induction and Classification*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 346-349.

AGIRRE, Eneko & RIGAU, German (1995). *A Proposal for Word Sense Disambiguation using Conceptual Distance*. In Proceedings of the International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

AGIRRE, Eneko & DE LACALLE, Oier Lopez (2003). *Clustering WordNet Word Senses*. In Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, pp. 121-130.

AGIRRE, Eneko & EDMONDS, Philip (2007). (eds.) *Word Sense Disambiguation: Algorithms and Applications*, Springer, Dordrecht, The Netherlands.

AGIRRE, Eneko, MARTINEZ, David, DE LACALLE, Oier Lopez & SOROA, Aitor (2006). *Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm*. In Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing, Association for Computational Linguistics, New York City, June, pp. 89-96.

AHRENBERG, Lars, ANDERSSON, Mikael & MERKEL, Magnus (1998). *A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL/COLING'98), Montreal, Canada, pp. 29-35.

ALSHAWI, Hiyan, BANGALORE, Srinivas & DOUGLAS, Shona. (1998). *Automatic Acquisition of Hierarchical Transduction Models for Machine Translation*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98), Vol. 1, Montreal, Canada, pp. 41-47.

ALTENBERG, Bength & GRANGER, Sylviane (2002a). (eds.) *Lexis in Contrast: Corpus-based approaches*. Amsterdam/Atlanta: John Benjamins Publishing Company.

ALTENBERG, Bength & GRANGER, Sylviane (2002b). *Recent trends in cross-linguistic lexical studies*, In ALTENBERG, Bength & GRANGER, Sylviane (eds.), pp. 3-48.

AMBROZIAK, Jacek & WOODS, William A. (1998). *Natural Language Technology in Precision Content Retrieval*. In Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA'98), August 18-21, Moncton, New Brunswick, Canada.

APIDIANAKI, Marianna (2005). *Translation prediction using word cooccurrence graphs*. In Proceedings of Corpus Linguistics, Birmingham, 14-17 July, Vol. 1(1).

APIDIANAKI, Marianna (2006). *Traitement de la polysémie lexicale dans un but de traduction*. In Proceedings of TALN'06, Leuven, Belgium, 10-13 April, Vol.1, pp. 53-62.

APIDIANAKI, Marianna (2007). *Repérage de sens et désambiguïsation dans un contexte bilingue*. In Proceedings of TALN'07, Toulouse, 5-8 June, Vol.1, pp. 207-216.

APIDIANAKI, Marianna (2008). *Translation-oriented sense induction based on parallel corpora*. In Proceedings of Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, 5-8 June. (to appear)

APRESJAN, Jurij (1973). *Regular polysemy*, Linguistics, 142, pp. 5-32.

ARDUINI, Stefano & HODGSON, Robert (eds.) (2004) *Similarity and difference in translation*. Proceedings of the International Conference on Similarity and Translation, New York, 31 May – 1 June 2001, Rimini: Guaraldi.

ARNOLD, Jack B. (1971). *A Multidimensional Scaling Study of Semantic Distance*, Journal of Experimental Psychology Monograph, 90(2), pp. 349-372.

ATKINS, B.T. Sue & ZAMPOLLI, Antonio (eds.) (1994). *Computational approaches to the lexicon*, Oxford University Press.

AUDIBERT, Laurent (2003). *Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences*. In Proceedings of TALN'03, Batz-sur-Mer, 11-14 June, pp. 35-44.

BAKER, Mona, FRANCIS, Gill & TOGNINI-BONELLI, Elena (eds.) (1993) *Text and Technology, In Honour of John Sinclair*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

BAKER, Mona (1993). *Corpus Linguistics and Translation Studies*. In BAKER, Mona, FRANCIS, Gill & TOGNINI-BONELLI, Elena (eds.), pp. 233-250.

BAKER, Mona (1995). *Corpora in Translation Studies: An overview and some suggestions for future* research. Target 7(2), pp. 223-243.

BAKER, Mona (1996). *Corpus-based translation studies: the challenges that lie ahead*. In SOMERS, Harold (ed.), pp. 175-186.

BAKER, Mona (1998). *Réexplorer la langue de la traduction: une approche par corpus*. *Meta*, 43(4), pp. 480-485.

BAKER Mona, FRANCIS, Gill & TOGNINI-BONELLI, Elena (eds.) (1993). *Text and Technology, In Honour of John Sinclair*, Philadelphia, Amsterdam: John Benjamins Publishing Company.

BALLY, Charles. (1940). *L'arbitraire du signe. Valeur et signification*, Le français moderne, pp. 193-206.

BANERJEE, Satanjeev & LAVIE, Alon (2005). *METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 43th Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, June, pp. 65-72.

BANERJEE, Satanjeev & PEDERSEN, Ted (2002). *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*. In Proceedings of CICLing, Mexico City, Mexico, pp. 136-145.

BANERJEE, Satanjeev & PEDERSEN, Ted (2003). *Extended gloss overlaps as a measure of semantic relatedness*. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August, pp. 805-810.

BANGALORE, Srinivas, RAMBOW, Owen & WHITTAKER, Steve (2000). *Evaluation Metrics for Generation*. In Proceedings of the First International Natural Language Generation Conference (INLG'00), Mitzpe Ramon, Israel, pp. 1-13.

BASILI, Roberto, DELLA ROCCA Michelangelo & PAZIENZA, Maria Teresa (1997). *Towards a Boostrapping Framework for Corpus Semantic Tagging*, In Proceedings of ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?", Washington, DC, April, pp. 66-73.

BECK, Nicolas (2006). *Application de méthodes de clustering traditionnelles et extension au cadre multicritère*, Mémoire de fin d'études, Université libre de Bruxelles.

BELLMAN RICHARD (1957). *Dynamic Programming*, Princeton University Press, Princeton, NJ, Republié par Dover Publications Inc., 2003.

BENTIVOGLI, Luisa, FORNER, Pamela & PIANTA, Emanuele (2004). *Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus*, In Proceedings of the

20th International Conference on Computational Linguistics (COLING'04), Geneva, Switzerland, 23-27 August, pp. 364-370.

BERLIN, Brent & KAY, Paul (1969). *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

BERTELS, Ann (2005). *A la découverte de la polysémie des spécificités du français technique*. In Proceedings of Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'05), pp. 575-584.

BERTELS Ann, SPEELMAN, Dirk & GEERAERTS, Dirk (2006). *Analyse quantitative et statistique de la sémantique dans un corpus technique*. In Proceedings of TALN'06, Leuven, Belgium, 10-13 April, pp. 73-82.

BILGER, Mireille (2002). *Corpus, méthodologie et applications linguistiques*, Paris : Honoré Champion et les Presses Universitaires de Perpignan.

BLACK, Ezra (1988). *An Experiment in Computational Discrimination of English Word Senses*. IBM Journal of Research and Development, Vol.32, pp. 185-194.

BOAS, Franz (ed.) (1911). *Introduction to Handbook of American Indian languages*. Bureau of American Ethnology Bulletin, 40, Washington, DC: Smithsonian Institution, pp. 1-83. Reprinted in a volume edited by HOLDER, Preston. (1966). Lincoln, NE: University of Nebraska Press (1991), pp. 1-79.

BORIN, Lars (ed.) (2002a). *Parallel corpora, Parallel Worlds: selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April 1999*. Amsterdam, New York: Rodopi.

BORIN, Lars (2002b). *Alignment and tagging*. In BORIN, Lars (ed.), pp. 207-218.

BOWKER, Lynne, CRONIN, Michael, KENNY, Dorothy, PEARSON, Jennifer (eds.) (1998) *Unity in Diversity? Current trends in Translation Studies*. Manchester: St Jerome Publishing.

BRANTS, Thorsten (2000). *TnT A Statistical Part-Of-Speech Tagger*. In Proceedings of the 6th Applied NLP Conference (ANLP'00), 29 April – 3 May, Seattle, WA.

BREAL, Michel (1899). *Essai de sémantique* (2e éd.). Paris, Librairie Hachette et Cie.

BRILL, Eric (1995). *Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging*. In Proceedings of the Third Workshop on Very Large Corpora, Cambridge MA, June, pp. 1-13.

BROWN Peter F., COCKE John, DELLA PIETRA Stephen A., DELLA PIETRA Vincent J., JELINEK Fredrick, MERCER, Robert L. & ROOSSIN, Paul S. (1988). *A statistical*

approach to language translation. In Proceedings of the 12th International Conference on Computational Linguistics (COLING'88), Budapest, Vol.1, pp.1-6.

BROWN Peter F., DELLA PIETRA Stephen A., DELLA PIETRA Vincent J. & MERCER, Robert L. (1990a). *A Statistical Approach to Machine Translation*. Computational Linguistics, 16(2), June, pp. 79-85.

BROWN Peter F., DELLA PIETRA Vincent J., DESOUZA, Peter V., LAI, Jennifer C. & MERCER, Robert L. (1990b). *Class-based n-gram models of natural language*. Computational Linguistics, 18(4), pp. 467-479.

BROWN, Peter F., LAI, Jennifer C. & MERCER, Robert L. (1991a). *Aligning sentences in parallel corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, California, pp. 169-176.

BROWN, Peter F., DELLA PIETRA, Stephen A., DELLA PIETRA, Vincent J. & MERCER, Robert L. (1991b). *Word-sense disambiguation using statistical methods*. In Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, California, pp. 264-270.

BROWN Peter F., DELLA PIETRA, Stephen A., DELLA PIETRA Vincent J. & MERCER, Robert L. (1993). *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics 19(2), pp. 263-311.

BRUCE, Rebecca & WIEBE, Janyce (1994a). *A new approach to word sense disambiguation*. In Proceedings of the ARPA Workshop on Human Language Technology, Princeton, New Jersey, March, pp. 236-241.

BRUCE, Rebecca & WIEBE, Janyce (1994b). *Word-Sense Disambiguation Using Decomposable Models*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruses, NM. (ACL'94), pp. 139-145.

BRUCE, Rebecca & WIEBE, Janyce (1998). *Word-Sense Distinguishability and Inter-Coder Agreement*. In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98), Association for Computational Linguistics SIGDAT, Granada, Spain, June 1998, pp. 53-60.

BRUN, Caroline, JACQUEMIN, Bernard & SEGOND, Frédérique (2001). *Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale*. Traitement Automatique des Langues (TAL), *Lexiques Sémantiques*, Volume 42 (3), pp. 667-690.

BUDANITSKY, Alexander & HIRST, Graeme (2001). *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. In Proceedings of the Workshop on WordNet and Other Lexical Resources: Second Meeting of the

North American Chapter of the Association for Computational Linguistics, Pittsburgh, pp. 29-34.

BUITELAAR, Paul (1998). *CORELEX : Systematic Polysemy and Underspecification*, PhD Thesis, Computer Science, February, Brandeis University.

BUITELAAR Paul, MAGNINI Bernardo, STRAPPARAVA Carlo & VOSSEN, Piek (2007). *Domain-Specific WSD*. In AGIRRE Eneko et EDMONDS, Philip (eds.), pp. 275-298.

BURGESS, Curt & LUND, Kevin (1997). *Modeling parsing constraints with high-dimensional context space*. Language and Cognitive Processes, 12(2-3), pp. 177-210.

CABEZAS, Clara & RESNIK, Philip (2005). *Using WSD Techniques for Lexical Selection in Statistical Machine Translation*. Rapport technique CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42, July.

CACCIARI, Cristina (ed.) (1995). *Similarity in Language, Thought and Perception*, Brussels: Brepols.

CALLISON-BURCH, Chris, KOEHN, Philipp & OSBORNE, Miles (2006a). *Improved Statistical Machine Translation Using Paraphrases*. In Proceedings of the Human Language Technology conference – North American chapter of the Association for Computational Linguistics (HLT-NAACL), New York, USA, pp. 17-24.

CALLISON-BURCH, Chris, OSBORNE, Miles & KOEHN, Philipp (2006b). *Re-evaluating the Role of BLEU in Machine Translation Research*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), pp. 249-256.

CALLISON-BURCH, Chris, FORDYCE, Cameron, KOEHN, Philipp, MONZ, Christof & SCHROEDER, Josh (2007). *(Meta-) Evaluation of Machine Translation*. In Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, June 23, pp. 136-158.

CALZOLARI, Nicoletta & CORAZZARI, Ornella (2000). *Senseval/Romanseval: The Framework for Italian*. Computers and the Humanities 34, pp. 61-78.

CARPUAT, Marine, WEIFENG, Su & WU, Dekai (2004). *Augmenting ensemble classification for Word Sense Disambiguation with a Kernel PCA model*. In Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, July, pp. 88-92.

CARPUAT, Marine & WU, Dekai (2005a). *Word Sense Disambiguation vs. Statistical Machine Translation*. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Ann Arbor, Michigan, June, pp. 387-394.

CARPUAT, Marine & WU, Dekai (2005b). *Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation*. In Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP'05). Jeju Island, Republic of Korea, October, pp. 122-127.

CARPUAT, Marine, SHEN, Yihai, XIAOFENG, Yu & WU, Dekai (2006). *Towards Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation*. In Proceedings of the International Workshop on Spoken Language Translation, 27-28 November, Kyoto, Japan, pp. 37-44.

CARPUAT, Marine & WU, Dekai (2007a). *Improving Statistical Machine Translation using Word Sense Disambiguation*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07), Prague, Czech Republic, June, pp. 61-72.

CARPUAT, Marine & WU, Dekai (2007b). *How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation*. In Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2007), 7-9 September, Skövde, Sweden, pp. 43-52.

CARROLL, John B. (ed.) (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge: MIT Press, 1956.

CATFORD, John C. (1965). *A Linguistic Theory of Translation*, London, Oxford University Press.

CHAN, Yee Seng & NG, Hwee Tou (2005). *Scaling Up Word Sense Disambiguation via Parallel Texts*. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI'05), Pittsburgh, Pennsylvania, pp. 1037-1042.

CHAN, Yee Seng, NG, Hwee Tou & CHIANG, David (2007). *Word Sense Disambiguation Improves Statistical Machine Translation*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic, 23-30 June, pp. 33-40.

CHESTERMAN, Andrew (ed.) (1989). *Readings in Translation Theory*, Oy Finn Lectura Ab, Finland.

CHESTERMAN, Andrew (1998). *Contrastive Functional Analysis*, Amsterdam / Philadelphia : John Benjamins Publishing Company.

CHESTERMAN, Andrew (2004). *Where Is Similarity?* In ARDUINI, Stefano & HODGSON, Robert (eds.), pp. 63-75.

CHIANG, David (2005). *A hierarchical phrase-based model for statistical machine translation*. In Proceedings of the 43th Annual meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, June, p. 263-270.

CHIBOUT, Karim, MARIANI, Joseph, MASSON, Nicolas & NEEL, Françoise (eds.) (1999). *Ressources et Evaluation en Ingénierie des Langues*, Editions Duculot, Aupelf-Uref.

CHKLOVSKI, Timothy, MIHALCEA, Rada, PEDERSEN, Ted & PURANDARE, Amruta (2004). *The Senseval-3 multilingual english-hindi lexical sample task*. In Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems, Association for Computational Linguistics, July, Barcelone, Spain, pp. 5-8.

CHODOROW, Martin S., BYRD Roy J. & HEIDORN, George E. (1985). *Extracting semantic hierarchies from a large on-line dictionary*. In Proceedings of the 23th Annual Meeting of the Association for Computational Linguistics (ACL'85), pp. 299-304.

CHOUEKA, Yaacov & LUSIGNAN, Serge (1985). *Disambiguation by Short Contexts*. Computers and the Humanities 19, pp. 147-157.

CHURCH, Kenneth W. & PATRICK, Hanks (1990). *Word association norms, mutual information and lexicography*. In Computational Linguistics 16(1), pp. 22-29.

CHURCH, Kenneth W., GALE, William A., HANKS, Patrick, HINDLE, Donald & MOON, Rosamund (1994). *Lexical substitutability*. In ATKINS, B.T.S. et ZAMPOLLI, A. (eds.), pp. 153-177.

COLLINS, Allan M. & LOFTUS, Elisabeth F. (1975). *A spreading activation theory of semantic memory*, Psychological Review, 82(6), pp. 407-428.

CONDAMINES, Anne & REBEYROLLE, Josette (1997). *Point de vue en langue spécialisée*, Meta XLII, 1, pp. 174-184.

CORNUEJOLS, Martine (2003). *Sens du mot, sens de l'image*. L'Harmattan, Paris.

COUGHLIN, Deborah (2003). *Correlating automated and human assessments of machine translation quality*. In Proceedings of the Machine Translation Summit IX, New Orleans, pp. 23-27.

COWIE, Jim, GUTHRIE, Joe & GUTHRIE, Louise (1992). *Lexical Disambiguation using Simulated Annealing*, In Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes, France, pp. 359-365.

CRESTAN, Eric, EL-BÈZE, Marc & DE LOUPY, Claude (2003). *Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique ?*. In Proceedings of TALN'03, Batz-sur-Mer, 11-14 June, pp.85-94.

CRUSE, Alan D. (1986). *Lexical Semantics*, Cambridge Textbooks in Linguistics, Cambridge, New York.

CRUSE, Alan D. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics*, Oxford University Press.

CURRAN, James Richard (2003). *From Distributional to Semantic Similarity*, PhD Thesis, University of Edinburgh.

CURRAN, James Richard & MOENS, Marc (2002). *Improvements in Automatic Thesaurus Extraction.* In Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, USA, 12 July, pp. 59-66.

DAGAN, Ido (1991). *Lexical disambiguation: sources of information and their statistical realization*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, California, 18-21 June, pp. 341-342.

DAGAN, Ido & ITAI, Alon (1994). *Word sense disambiguation using a second language monolingual corpus*. Computational Linguistics 20(4), pp. 563-596.

DAGAN Ido, ITAI, Alon & SCHWALL, Ulrike (1991). *Two languages are more informative than one*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, California, 18-21 June, pp. 130-137.

DAGAN, Ido, MARCUS, Shaul & MARKOVITCH, Shaul (1993). *Contextual word similarity and estimation from sparse data*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), Columbus, Ohio, 22-27 June, pp. 164-171.

DAGAN Ido, PEREIRA, Fernando & LEE, Lillian (1994). *Similarity-based estimation of word cooccurrence probabilities*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), Las Cruces, New Mexico, 27-30 June, pp. 272-278.

DANIELSSON, Pernilla & RIDINGS, Daniel (1997). *Practical Presentation of a « Vanilla » Aligner*. In Proceedings of the TELRI Workshop on Alignment and Exploitation of Texts, Institute Jožef Stefan, Ljubljana.

DE SAUSSURE, Ferdinand (1972). *Cours de linguistique générale*, Publié par Charles Bally et Albert Séchehaye, Avec la collaboration de Albert Riedlinger, Edition critique préparée par Tullio de Mauro, Éditions Payot, Paris.

DEBILI, Fathi & SAMMOUDA, Elyes (1992). *Aligning Sentences in Bilingual Texts, French-English and French-Arabic*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING'92), Nantes, 23-28 August, pp. 517-524.

DEERWESTER, Scott, DUMAIS, Susan T., FURNAS, George W., LANDAUER, Thomas K. & HARSHMAN, Richard (1990). *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6), pp. 391-407.

DEMPSTER, Arthur, LAIRD, Nan & RUBIN, Donald (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Series B (Methodological), 39(1), pp. 1-38.

DES TOMBES, Louis (1992). *Is Translation Symmetric?*. Meta Vol. 37(4), pp. 791-801.

DESCAMPS Jean-Luc, MOCHET, M.A., LEWIN, T. LAMIZET Bernard, & COSTES, D. (1992). *Sémantique et concordances*, Collection « Saint-Cloud », Paris : Klincksieck.

DIAB, Mona (2003). *Word Sense Disambiguation within a Multilingual Framework*. PhD Thesis, University of Maryland.

DIAB, Mona & RESNIK, Philip (2002). *An Unsupervised Method for Word Sense Tagging using Parallel Corpora*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, 6-12 July, pp. 255-262.

DIAB, Mona & FINCH, Steven (2000). *A statistical word level translation model for comparable corpora*. In Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO'00), Paris, France.

DICKENS, Alison & SALKIE, Raphael (1996). *Comparing Bilingual Dictionaries with a Parallel Corpus,* In Proceedings of EURALEX'96, GELLERSTAM *et al*. (eds.), Göteborg University, pp. 551-559.

DIMARCO, Chrysanne, HIRST, Graeme et SSTEDE, Manfred (1993). *The semantic and stylistic differentiation of synonyms and near-synonyms*. In Proceedings of the AAAI Spring Symposium on Building Lexicons for Machine Translation, mars 1993, Stanford, CA, pp. 114-121.

DIRVEN, Rene & FRIED, Vilem (eds.) (1987). *Functionalism in Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

DODDINGTON, George (2002). *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. In Proceedings of the 2nd International conference on human language technology research, 24-27 March, San Diego,

California, Mitchell Marcus (ed.), San Francisco, CA: Morgan Kaufmann for DARPA, pp. 138-145.

DOLAN, William B. (1994). *Word Sense Ambiguation: Clustering Related Senses*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Morristown, New Jersey, pp. 712-716.

DOLLERUP, Cay & LODDEGAARD, Anne (eds.) (1991). *Teaching Translation and Interpreting, Training, Talent and Experience*, Papers from the First Language International Conference, Elsinore, Denmark, John Benjamins Publishing Company, Amsterdam/Philadelphia.

DONG, Zhen Dong (1998). *Knowledge Description: What, How and Who?*. In Proceedings of International Symposium on Electronic Dictionary. Tokyo, Japan.

DOROW Beate & WIDDOWS, Dominic (2003). *Discovering Corpus-Specific Word Senses*. In Proceedings of the Conference of the European chapter of the Association for Computational Linguistics (EACL'03) (Conference Companion, research notes and demos), Budapest, Hungary, pp. 79-82.

DUFF, Alan (1981). *The Third Language: Recurrent problems of translation into English*, Pergamon Press: Oxford, New York, Toronto.

DUFOUR, Nicolas (1997). *DEFIDIC, a lexical database for computerized translation selection*. RISHH vol. 33, Liège, pp. 79-111.

DUNNING, Ted (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19(1), pp. 61-74.

DYMETMAN, Marc & ISABELLE, Pierre (1988). *Reversible logic grammars for machine translation*. In Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 12-14, Carnegie-Mellon University, Pittsburgh, PA.

DYVIK, Helge (1998a). *Translations as semantic mirrors: From Parallel Corpus to WordNet*. In Proceedings of the W13 Workshop: Multilinguality in the lexicon II, Brighton, UK, The 13th biennial European Conference on Artificial Intelligence ECAI'98, pp. 24-44.

DYVIK, Helge (1998b). *A translational basis for semantics*, In JOHANSSON, Stig et OKSEFJELL, Signe (eds.), pp. 51-86.

DYVIK, Helge (2003). *Translations as a Semantic Knowledge Source*, Draft, (http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transknow.pdf).

DYVIK, Helge (2005). *Translations as a semantic knowledge source*. In Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn: Institute of Cybernetics at Tallinn University of Technology, Institute of the Estonian Language.

EBELING, Jarle (1998). *Contrastive linguistics, translation, and parallel corpora*, META 43(4), pp. 602-615.

EDMONDS, Philip (1997). *Choosing the Word Most Typical in Context Using a Lexical Co-occurrence Network*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), 7-12 July, Madrid, Spain, pp. 507-509.

EDMONDS, Philip (1999). *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*, PhD Thesis, University of Toronto.

EDMONDS Philip (2002). *SENSEVAL: The evaluation of word sense disambiguation systems*, ELRA Newsletter, Vol. 7, No 3.

EDMONDS, Philip & KILGARRIFF, Adam (2002). *Introduction to the special issue on evaluating word sense disambiguation systems*. Natural Language Engineering 8(4), Cambridge University Press, pp. 279-291.

EDMONDS, Philip & GRAEME, Hirst (2002). *Near-Synonymy and Lexical Choice*. Computational Linguistics 28(2), pp. 105-144.

EDMONDS, Philip & COTTON, Scott (2002). *SENSEVAL-2: Overview*. In Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse, pp. 1-6.

ERTOZ, Leven, STEINBACH, Michael & KUMAR, Vipin (2001). *Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach*, In Proceedings of Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining, pp. 83-104.

FABRE, Cécile, HABERT, Benoit & LABBE, Dominique (1997). *La polysémie dans la langue générale et les discours spécialisés*. Sémiotiques, Numéro 13, December, pp. 15-30.

FANO, Robert M. (1961). *Transmission of Information*, MIT Press, Cambridge, Massachusetts.

FELLBAUM, Christiane (ed.) (1998). *WordNet. An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

FERRET, Olivier (2004a). *Discovering word senses from a network of lexical cooccurrences*. In Proceedings of the 20th International Conference on

Computational Linguistics (COLING'04), 23-27 August, Geneva, Switzerland, pp. 1326-1332.

FERRET, Olivier (2004b). *Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales*. In Proceedings of TALN'04, Fès, Maroc, 19-22 April.

FIRTH, John R. (1957a). *Papers in Linguistics, 1934-1951*, London/New York: Oxford University Press.

FIRTH, John R. (1957b). *A Synopsis of Linguistic Theory, 1930-1955*. In *Studies in Linguistic Analysis*, Special Volume of the Philological Society, Oxford: Basil Blackwell, 1962.

FIRTH, John R. (1968). *Linguistics and Translation*. In PALMER, Frank R., *Selected Papers of J.R. Firth: 1952-59*, London: Harlow: Longmans, pp. 84-95.

FRASER, Alexander & MARCU, Daniel (2006). *Semi-Supervised Training for Statistical Word Alignment*. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06), Sydney, Australia, pp. 769-776.

FRAWLEY, William (ed.) (1984a). *Translation: Literary, Linguistic, and Philosophical Perspectives*, Newark: University of Delaware Press, London and Toronto: Associated University Press.

FRAWLEY, William (ed.) (1984b). *Prolegomenon to a Theory of Translation*. In Frawley (1984a), pp. 159-175.

FRAWLEY, William (1992). *Linguistic Semantics*, Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

FREGE, Gottlob (1892). *Sens et dénotation*. In *Ecrits logiques et philosophiques*, Paris : Editions du Seuil, 1971, pp. 102-126.

FRENCH, Robert M. & LABIOUSEN, Christophe (2002). *Four problems with extracting human semantics from large text corpora*. In Proceedings of the 24th Annual Conference of the Cognitive Science Society, NJ: L.E.A.

FUCHS, Catherine (1985a). (ed.) *Aspects de l'ambiguïté et de la paraphrase dans les langues naturelles*, Editions Peter Lang, Berne.

FUCHS, Catherine (1985b). *Introduction : L'ambiguïté et la paraphrase, propriétés fondamentales des langues naturelles,* In FUCHS Catherine (ed.), pp. 7-35.

FUCHS, Catherine (1994). *Paraphrase et énonciation*. Paris : Editions Ophrys.

FUCHS, Catherine (1996). *Les ambiguïtés du français*. Paris : Editions Ophrys.

FUCHS, Catherine (1997). *Diversité des représentations linguistiques : quels enjeux pour la cognition ?*. In FUCHS, Catherine & ROBERT, Stéphane (eds.), pp. 5-24.

FUCHS, Catherine & ROBERT, Stéphane (eds.) (1997). *Diversité des Langues et Représentations Cognitives*. Paris : Editions Ophrys.

GALE, William A., & CHURCH, Kenneth W. (1991). *A program for aligning sentences in bilingual corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), 18-21 June, Berkeley, California, pp. 177-184.

GALE William A., CHURCH, Kenneth W. & YAROWSKY, David (1992a). *Using bilingual materials to develop word sense disambiguation methods*. In Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Transaltion, pp. 101-112.

GALE William A., CHURCH, Kenneth W. & YAROWSKY, David (1992b). *One Sense Per Discourse*. In Proceedings of the 4th DARPA Speech and Natural Language Workshop, pp. 233-237.

GALE William A., & CHURCH, Kenneth W. (1993). *A program for aligning sentences in bilingual corpora*. Computational Linguistics, 19(1), pp. 75-102.

GALE William A., CHURCH, Kenneth W. & YAROWSKY, David (1993). *A Method for Disambiguating Word Senses in a Large Corpus*. Computers and the Humanities, 26, pp. 415-439.

GARZA-CUARÓN, Beatriz (1991). *Connotation and Meaning*, Berlin, New York: Mouton de Gruyter.

GASPERIN, Caroline, GAMALLO, Pablo AGUSTINI, Alexandre, LOPES, Gabriel & DE LIMA, Vera (1991). *Using Syntactic Contexts for Measuring Word Similarity*, Workshop on Knowledge Acquisition & Categorisation, ESSLLI.

GAUSSIER, Eric (1998). *Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL'98), Montreal, Canada, pp. 444-450.

GAUSSIER, Eric & LANGE, JEAN-MARC (1995). *Modèles statistiques pour l'extraction de lexiques bilingues*. Traitement Automatique des Langues (TAL), 36(1-2), pp. 133-155.

GAVRILIDOU, Maria, LABROPOULOU, Peny, DESIPRI, Elina, GIOULI, Voula, ANTONOPOULOS, Vasilis & PIPERIDIS, Stelios (2004). *Building parallel corpora for*

eContent professionals. In Proceedings of MLR 2004, PostCOLING Workshop on Multilingual Linguistic Resources, August 28, Geneva, Switzerland.

GEERAERTS, Dirk (1993). *Vagueness's puzzles, polysemy's vagaries*. In Cognitive Linguistics 4(3), pp. 223-272.

GELLERSTAM, Martin, JARBORG Jerker, MALMGREN, Sven Göran, NOREN, Kerstin, ROGSTROM, Lena & PAPMEHL, Catarina Roder (eds.) (1993). In Proceedings of *Euralex'96*, Göteborg : Göteborg University, Sweden.

GOLAN, Igal, LAPPIN, Shalom & RIMON, Mori (1988). *An Active Bilingual Lexicon for Machine Translation*. In Proceedings of the 12th International Conference on Computational Linguistics (COLING'88), Budapest, Hungary, Vol. 1, pp. 205-211.

GOLDSTONE, Robert L., MEDIN, Douglas L. & HALBERSTADT, Jamin (1997). *Similarity in context*. Memory & Cognition, 25(2), pp. 237-255.

GONZALO, Julio, IRINA CHUGUR, Irina & VERDEJO, Felisa (2000). *Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index*. In Proceedings of the ACL Workshop on Word Senses and Multi-linguality, pp. 10-18.

GOODMAN, Nelson (1952). *On Likeness of Meaning*. In LINSKY Leonard (ed.), pp. 67-74.

GORFEIN, David S. (ed.) (1989). *Resolving Semantic Ambiguity*, New York/Berlin/London: Springer-Verlag.

GRABAR, Natalia & ZWEIGENBAUM, Pierre (2005). *Adaptation de synonymes de la langue générale pour le traitement automatique des termes médicaux*. In Proceedings of Journées francophones d'informatique médicale, Lille, 2005.

GRANGER, Sylviane, LEROT, Jacques & PETCH-TYSON, Stephanie (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam, New York: Rodopi.

GREFENSTETTE, Gregory (1992). *SEXTANT: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis*. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92), 28 June – 2 July, Newark, Delaware, pp. 324-326.

GREFENSTETTE, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*. Boston/Dordrecht/London: Kluwer Academic Publishers.

GREIMAS, Algirdas Julien (1986). *Sémantique Structurale*. Paris : Presses Universitaires de France.

GUILBERT Louis (1973). *La spécificité du terme scientifique et technique*. Langue française, 17, pp. 5-17.

GUTHRIE Joe A., GUTHRIE Louise, WILKS Yorick & AIDINEJAD Homa (1991). *Subject-Dependent Co-Occurrence and Word Sense Disambiguation*, In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, California, 18-21 June, pp. 146-152.

HAAS, William. (1968). *The theory of translation*. In PARKINSON, G. H. R. (ed.), *The theory of meaning*, London: Oxford University Press, pp. 86-108.

HABERT, Benoît, NAZARENKO, Adeline & SALEM, André (1997). *Les linguistiques de corpus*. Paris : Armand Colin/Masson.

HALLIDAY, Michael Alexander Kirkwood & HASAN, Ruqaiya (1976). *Cohesion in* English, London: Longman.

HARABAGIU, Sanda M., MILLER, George, A. & MOLDOVAN, Dan I. (1999). *WordNet 2 – A Morphologically and Semantically Enhanced Resource*. In Proceedings of SIGLEX (Special Interest Group on the Lexicon), June, University of Maryland.

HARDWICK, Lorna (2000). *Translating Words, Translating Cultures*, Gerald Duckworth and Co Ltd., London.

HARDIN C. L. & MAFFI L. (eds.) (1997). *Color categories in language and thought*. Cambridge: Cambridge University Press.

HARRIS, Brian (1988a). *Are you bi-textual*?. Language Technology, 7, pp. 41-41.

HARRIS, Brian (1988b). *Bi-texts: A new concept in translation theory*. Language Monthly, 54, pp. 8-10.

HARRIS, Zellig (1954). *Distributional structure*. Word 10: pp. 146-162. Reprinted in KATZ Jerold J. (ed.), pp. 26-47.

HARRIS, Zellig (1968). *Mathematical Structures of Language*. New York: Wiley.

HAYES, Philip J. (1977). *On semantic nets, frames and associations*. In Proceedings of the 5th International Joint Conference on Artificial Intelligence, Cambridge, MA, pp. 99-107.

HEARNE, Mary & WAY, Andy (2006). *Disambiguation Strategies for Data-Oriented Translation*. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT-2006), 19-20 June, Oslo, Norway, pp. 59-68.

HEARST, Marti A. (1991). *Noun Homograph Disambiguation Using Local Context in Large Text Corpora*. In Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research, Oxford, October.

HINDLE, Donald (1990). *Noun classification from predicate-argument structures*. In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburgh, Pennsylvania, pp. 268-275.

HIRST, Graeme (1987). *Semantic interpretation and the resolution of ambiguity*, Cambridge, London, New York: Cambridge University Press.

HIRST, Graeme & ST-ONGE, David (1998). *Lexical chains as representations of context for the detection and correction of malapropisms*. In FELLBAUM, Christiane (ed.), pp. 305-332.

HOCKEY, Susan & IDE, Nancy (eds.), Research in Humanities Computing 2, Oxford: Oxford University Press.

HOVY, Eduard, MITCHELL, Marcus, PALMER, Martha, RAMSHAW, Lance & WEISCHEDEL, Ralph (2006). *Ontonotes: The 90% solution*. In Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL), New York, June, pp. 57-60.

HUTCHINS, John W. & SOMMERS, Harold L. (1992). *Introduction to Machine Translation*. London: Academic Press.

IDE, Nancy & VERONIS, Jean (1990). *Mapping Dictionaries: A Spreading Activation Approach*. In Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary, Waterloo, Canada, pp. 52-64.

IDE, Nancy & VERONIS, Jean (1993). *Refining Taxonomies Extracted from Machine Readable Dictionaries*. In HOCKEY, Susan & IDE, Nancy (eds.), pp. 145-170.

IDE, Nancy & VERONIS, Jean (1998). *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art*. Computational Linguistics, 1998, vol. 24(1), March, pp. 1-40.

IDE, Nancy (1999a). *Cross-lingual Sense Determination: Can it work?* Computers and the Humanities: 34(1-2), pp. 223-234.

IDE, Nancy (1999b). *Parallel Translations as Sense Discriminators*. In Proceedings of SIGLEX (Special Interest Group on the Lexicon): Standardizing Lexical Resources, ACL'99 Workshop, College Park, Maryland, pp. 52-61.

IDE, Nancy, ERJAVEC, Tomaz & TUFIS, Dan (2001). *Automatic sense tagging using parallel corpora*. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, 27-30 November, Tokyo, Japan, pp. 83-89.

IDE, Nancy, ERJAVEC, Tomaz & TUFIS, Dan (2002). *Sense discrimination with parallel corpora*. In Proceedings of the ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, 11 July, Philadelphia, pp. 54-60.

IDE, Nancy & WILKS, Yorick (2007). *Making Sense About Sense*. In AGIRRE Eneko et EDMONDS, Philip (eds.), pp. 47-73.

ION, Radu & TUFIS, Dan (2004). *Multilingual Word Sense Disambiguation Using Aligned Wordnets*. Romanian Journal of Information Science and Technology, Vol. 7 (1-2), pp. 183-200.

ISABELLE, Pierre (1989). *Towards reversible MT systems*. In Proceedings of the 2nd International Machine Translation Summit (MT Summit), 16-18 August, Munich, pp. 67-68.

ISABELLE, Pierre (1992). *Bi-Textual Aids for Translators*. In Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Canada.

ISABELLE, Pierre, DYMETMAN, Marc, FOSTER, George, JUTRAS, Jean-Marc, MACKLOVITCH Elliott, PERRAULT François, REN, Xiaobo & SIMARD, Michel (1993). *Translation analysis and translation automation*. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93), 14-16 July, Kyoto, Japan, pp. 201-217.

IVIR, Vladimir (1987). *Functionalism in contrastive analysis and translation studies*. In DIRVEN, René & FRIED, Vilém (eds.), pp. 471-481.

JACQUEMIN, Christian (2001). *Spotting and Discovering Terms through Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

JAIN ANIL K., MURTY, Narasimha M. & FLYNN, Patrick J. (1999). *Data Clustering: A Review*, ACM Computing Surveys, Vol. 31 (3), September, pp. 264-323.

JASZCZOLT Katarzyna M. & TURNER, Ken (eds.) (2003). *Meaning Through Language Contrast*, Vol. 2, Amsterdam/Philadelphia: John Benjamins Publishing Company.

JENSEN, Robert E. (1969). *A Dynamic Programming Algorithm for Cluster Analysis*, Operations Research Society of America, 17, pp. 1034-1057.

JI, Hyungsuk & PLOUX, Sabine (2003). *Automatic Contexonym Organizing Model (ACOM)*. In Proceedings of the 25th Annual Conference of the Cognitive Science Society, pp. 622-627.

JI, Hyungsuk, PLOUX, Sabine & WEHRLI, Eric (2003). *Lexical Knowledge Representation with Contexonyms*. In Proceedings of the 9th Machine Translation Summit (MT Summit), September, New Orleans, pp. 194-201.

JIANG, Jay J. & CONRATH, David W. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING), Taipei, Taiwan, pp. 19-33.

JIN, Peng, WU, Yunfang & YU, Shiwen (2007). *SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 19-23.

JOHANSSON, Stig (1980). *The LOB corpus of British English texts: presentation and comments*. ALLC Journal, 1(1), pp. 25-36.

JOHANSSON, Stig (1998). *On the role of corpora in cross-linguistic research*. In JOHANSSON, Stig & OKSEFJELL, Signe (eds.), pp. 3-24.

JOHANSSON, Stig & HOFLAND, Knut (1987). *The Tagged Lob Corpus: Description and Analyses*. In Corpus Linguistics and Beyond, Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora, pp. 1-20.

JOHANSSON, Stig & HOFLAND, Knut (1994). *Towards an English-Norwegian parallel corpus*. In *Creating and Using English Language Corpora*, Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, FRIES, Udo, TOTTIE, Gunnel & SCHNEIDER, Peter (eds.), Zürich 1993, Amsterdam: Rodopi, pp. 25-37.

JOHANSSON, Stig & OKSEFJELL, Signe (eds.) (1998). *Corpora and Cross-Linguistic Research: Theory, Method and Case Studies*. Amsterdam/Atlanta: Rodopi.

JORGENSEN, Julia C. (1990). *The Psychological Reality of Word Senses*. Journal of Psycholinguistic Research, Vol. 19 (3), pp. 167-190.

JUMPELT, R. J. (1961). *On the objectivizability of translation*, Epilogue de *Die Übersetzung naturwissenschaftlicher und technischer Literatur*. Reprinted In CHESTERMAN, Andrew (ed.), 1989, pp. 33-36.

JUSTESON, John S. & KATZ, Slava M. (1991). *Co-occurrences of Antonymous Adjectives and Their Contexts*. Computational Linguistics, Vol. 17(1), pp. 1-19.

KAJI, Hiroyuki & MORIMOTO, Yasutsugu (2002). *Unsupervised word sense disambiguation using bilingual comparable corpora*. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Taipei, Taiwan, pp. 1-7.

KAJI, Hiroyuki (2003). *Word Sense Acquisition from Bilingual Comparable Corpora*. In Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL), Edmonton, May-June 2003, pp. 32-39.

KAPLAN, Abraham (1955). *An experimental study of ambiguity and context*, Mechanical Translation, 2(2), November, pp. 39-46.

KATZ Jerold J. (ed.) (1985) *The Philosophy of Linguistics*. Oxford Readings in Philosophy. New York: Oxford University Press.

KAY, Martin (1980). *The Proper Place of Men and Machines in Language Translation*. Research Report CSL-80-11, Xerox Palo Alto Research Center, Reprinted in NIRENBURG *et al*. (eds.) (2003), pp. 212-232.

KAY, Martin & ROSCHEISEN, Martin (1988). *Text-Translation Alignment*. Technical Report, Xerox Palo Alto Research Center. Published in Computational Linguistics, 19(1), 1993, pp. 121-142.

KAY, Paul, BERLIN, Brent, MAFFI, Luisa & MERRIFIELD, William (1997). *Color naming across languages.* In HARDIN C. L. et MAFFI L. (eds.), pp. 21-56.

KAYSER, Daniel & COULON, Daniel (1981). *Variable-Depth Natural Language Understanding*. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada, August, pp. 64-66.

KENNY DOROTHY (2001). *Lexis and Creativity in Translation: A corpus-based study*. Manchester: St Jerome Publishing.

KIKUI, Genichiro (1999). *Resolving Translation Ambiguity using Non-parallel Bilingual Corpora*. In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, pp. 31-36.

KILGARRIFF, Adam (1993). *Dictionary word sense distinctions: An inquiry into their nature*. Computers and the Humanities 26, pp. 365-387.

KILGARRIFF, Adam (1997a). *Putting Frequencies in the Dictionary*. International Journal of Lexicography, 10(2), pp. 135-155.

KILGARRIFF, Adam (1997b). *Evaluating Word Sense Disambiguation Programs: Progress report*. In Proceedings of the SALT workshop on Evaluation in Speech and Language Technology, Sheffield, UK.

KILGARRIFF, Adam (1997c). *I don't believe in word senses*. Computers and the Humanities 31(2), pp. 91-113.

KILGARRIFF, Adam (1998a). *SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs*. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Granada, May, pp. 581-588.

KILGARRIFF, Adam (1998b). *Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs*. Computer Speech and Language 12(4), Special Issue on Evaluation, pp. 453-472.

KILGARRIFF, Adam (2001). *English lexical sample task description*. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse, France.

KILGARRIFF, Adam (2007). *Word Senses*. In AGIRRE, Eneko & EDMONDS, Philip (eds.), pp. 29-46.

KILGARRIFF, Adam & ROSENZWEIG, Joseph (2000a). *Framework and results for English SENSEVAL*. Computers and the Humanities, 34(1-2), Special Issue on SENSEVAL, KILGARRIFF, Adam & PALMER, Martha (eds.), pp. 15-48.

KILGARRIFF, Adam & ROSENZWEIG, Joseph (2000b). *English SENSEVAL: Report and Results*. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, May-June, pp. 1239-1244.

KITAMURA, Mihoko & MATSUMOTO, Yuji (1997). *Automatic Extraction of Word Sequence Correspondences in Parallel Corpora*. In Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark, pp. 79-87.

KLAPAFTIS, Ioannis P. & MANANDHAR, Suresh (2007). *UOY: A Hypergraph Model For Word Sense Induction et Disambiguation*. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 414-417.

KLEIBER, Georges (1990). *La sémantique du prototype : catégories et sens lexical*, Presses Universitaires de France, Paris.

KLEIBER, Georges (1999). *Problèmes de sémantique: la polysémie en questions*. Presses Universitaires du Septentrion, Paris.

KLEIN, Devorah E. & MURPHY, Gregory L. (2001). *The Representation of Polysemous Words*. Journal of Memory and Language, 45, pp. 259-282.

KOEHN, Philipp (2003). *Europarl: a Multilingual Corpus for Evaluation of Machine Translation*, Draft.

KOEHN, Philipp (2004). *Pharaoh: A beam search decoder for phrase-based statistical machine translation models*. In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas (AMTA'04), Washington DC, September 28-October 2, pp. 115-124.

KOEHN, Philipp (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*. In Proceedings of 10th MT Summit, Phuket, Thailand, pp. 79-86.

KONTOSTATHIS, April & POTTENGER, William M. (2002). *Detecting Patterns in the LSI Term-Term Matrix.* Workshop on the Foundation of Data Mining and Discovery, *IEEE International Conference on Data Mining* (ICDM'02), December, Terrsa, Maebashi, Japan.

KROVETZ, Robert & CROFT, William Bruce (1989). *Word Sense Disambiguation Using Machine-Readable Dictionaries*. In Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, SIGIR'89, Cambridge, MA, pp. 127-136.

KROVETZ, Robert & CROFT, William Bruce (1992). *Lexical Ambiguity and Information Retrieval*. ACM Transactions on Information Systems, 10:2, pp. 115-141.

KUCERA, Henry & FRANCIS, W. Nelson (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

LABROPOULOU PENNY, ELENA MANTZARI & MARIA GAVRILIDOU (1996). *Lexicon - Morphosyntactic Specifications: Language Specific Instantiation (Greek)*, PP-PAROLE, MLAP 63-386 report.

LANDAUER, Thomas K. & DUMAIS, Susan T. (1997). *A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge*. Psychological Review, 104(2), pp. 211-240.

LANDAUER, Thomas K., FOLTZ, Peter W. & LAHAM, Darrell (1998). *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, pp. 259-284.

LANDES, Shari, LEACOCK, Claudia & TENGI, Randee I. (1998). *Building Semantic Concordances*. In FELLBAUM, Christiane (ed.), pp. 199-216.

LANGLAIS, Philippe & EL-BEZE, Marc (1997). *Alignement de corpus bilingues : algorithmes et évaluation*. In CHIBOUT, Karim, MARIANI, Joseph, MASSON, Nicolas & NEEL, Françoise (eds.), pp. 127-142.

LARSON, Mildred L. (1984). *Meaning-based translation: A Guide to Cross-Language Equivalence*, University Press of America.

LAVIE, Alon, SAGAE, Kenji & JAYARAMAN, Shyamsundar (2004). *The Significance of Recall in Automatic Metrics for MT Evaluation*. In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas (AMTA'04), Washington DC, September 28-October 2, pp. 134-143.

LAVIE, Alon & AGARWAL, Abhaya (2007). *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*. In Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic, 23 June, pp. 228-231.

LAVIOSA, Sara (2002). *Corpus-based translation studies: Theory, Findings and Applications*. Amsterdam, New York: Rodopi.

LAVIOSA, Sara (2003). *Corpora and Translation Studies*. In GRANGER, Sylviane, LEROT, Jacques & PETCH-TYSON, Stephanie (eds.), pp. 45-54.

LEACOCK, Claudia, GEOFFREY, Towell & VOORHEES, Ellen M. (1993). *Corpus-based Statistical Sense Resolution*. In Proceedings of the ARPA Human Language Technology Workshop, San Francisco, Morgan Kaufman, pp. 260-265.

LEACOCK, Claudia, GEOFFREY, Towell & VOORHEES, Ellen M. (1996). *Towards building contextual representations of word senses using statistical models.* In BOGURAEV, Branimir & PUSTEJOVSKY, James (eds.), Corpus Processing for Lexical Acquisition. Cambridge, MA: MIT Press, pp. 97-113.

LEACOCK, Claudia & CHODOROW, Martin (1998). *Combining local context and WordNet similarity for word sense identification*. In FELLBAUM, Christiane, pp. 265-283.

LEACOCK, Claudia, CHODOROW, Martin & MILLER, George A. (1998). *Using Corpus Statistics and WordNet Relations for Sense Identification*. Computational Linguistics, 24(1), March, pp. 147-166.

LEE, Joon Ho, KIM, Myoung Ho & LEE, Yoon Joon (1989). *Information retrieval based on conceptual distance in IS-A hierarchies*. Journal of Documentation, 49(2), pp. 188-207.

LEMAIRE, Benoît & DENHIÈRE, Guy (2006). *Effects of High-Order Co-occurrences on Word Semantic Similarity*. Current Psychology Letters, Behaviour, Brain & Cognition, 18(1).

LE NY, Jean-François (1989). *Science cognitive et compréhension du langage*. Presses Universitaires de France, Paris.

LEPAGE, Yves. (2004). *Lower and higher estimates of the number of true analogies contained in a large multilingual corpus*. In Proceedings of the 20th International

Conference on Computational Linguistics (COLING'04), 23-27 August, Geneva, Switzerland, pp. 736-742.

LERAT, Pierre (1995). *Les langues spécialisées.* Paris : Presses Universitaires de France.

LESK, Michael (1986). *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone*. In Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June, pp. 24-26.

LEVÝ, Jiří (1967). *Translation as a decision process.* In *To honor Roman Jakobson*, Vol.2, 1171-1182. The Hague: Mouton. Reprinted in CHESTERMAN, Andrew (ed.) (1989), pp. 37-52.

LIN, Dekang (1998a). *An Information-Theoretic Definition of Similarity*. In Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, 1998, pp. 296-304.

LIN, Dekang (1998b). *Automatic retrieval and clustering of similar words*. In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'98), 10-14 August, Montreal, Canada, pp. 768-774.

LIN, Dekang, ZHAO, Shaojun, QIN, Lijuan & ZHOU, Ming (2003). *Identifying Synonyms among Distributionally Similar Words*. In Proceedings of the Eighteenth International Joint Conferences on Artificial Intelligence (IJCAI-03), 9-15 August, Acapulco, Mexico, pp.1492-1493.

LINSKY, Leonard (ed.) (1952). *Semantics and the Philosophy of Language*. University of Illinois Press, Urbana and Chicago.

LUCY, John Arthur (1992). *Language diversity and thought: a reformulation of the linguistic relativity hypothesis*. Cambridge University Press, Cambridge.

LUND, Kevin & BURGESS, Curt (1996). *Producing high-dimensional semantic spaces from lexical co-occurrence*. Behavior Research, Methods, Instruments and Computers, 28(2), pp.203-208.

LYONS, John (1968). *Introduction to theoretical linguistics*. Cambridge University Press.

LYONS, John (1990). *Sémantique linguistique*. Paris: Larousse.

LYSE, Gunn Inger (2006). *"Making sense of translations": Translation-based lexical information for Word Sense Disambiguation (WSD).* In VOLD, Eva Thue, LYSE, Gunn Inger & GJESDAL, Anje Müller (eds.), *New Voices in Linguistics*. Cambridge Scholars Press Ltd.

MACKLOVITCH, Elliott (1991). *Evaluating Commercial MT Systems*. In Proceedings of the Evaluators' Forum, 21-24 April, Les Rasses, Vaud, Switzerland, pp. 37-49.

MACKLOVITCH, Elliott, SIMARD, Michel & LANGLAIS, Philippe (2000). *TransSearch: A Free Translation Memory on the World Wide Web*. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, June, pp. 1201-1208.

MAGNINI, Bernardo & CAVAGLIA, Gabriela (2000). *Integrating subject field codes into WordNet*. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, June, pp. 1413-1418.

MAGNINI, Bernardo, STRAPPARAVA, Carlo, PEZZULO Giovanni & GLIOZZO, Alfio (2001). *Using Domain Information for Word Sense Disambiguation*. In Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France.

MALMKJÆR, Kirsten (1998). *Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators?* META, 43(4), pp. 534-541.

MALT, Barbara C., SLOMAN, Steven A. & GENNARI, Silvia P. (2003). *Universality and language specificity in object naming*. Journal of Memory and Language 49, Elsevier Science, USA, pp. 20-42.

MANNING, Christopher D. & SCHUTZE, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

MARCUS, Mitchell P., SANTORINI, Beatrice & MARCINKIEWICZ, Mary Ann (1993). *Building a large annotated corpus of English: The Penn Treebank*. Computational Linguistics, 19(2), pp. 313-330.

MARKMAN, Arthur B. & GENTNER, Dedre (1993). *Splitting the Differences: A Structural Alignment View of Similarity*. Journal of Memory and Language 32, pp. 517-535.

MARKOWITZ, Judith, AHLSWEDE, Thomas & EVENS, Martha (1986). *Semantically significant patterns in dictionary definitions*, In Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL'86), pp. 112-119.

MARQUEZ, Lluís, ESCUDERO, Gerard, MARTINEZ, David & RIGAU, German (2007). *Supervised Corpus-Based Methods for WSD*. In AGIRRE Eneko & EDMONDS, Philip (eds.), pp. 167-216.

MARRAFA, Palmira & RIBEIRO, Antonio (2001). *Quantitative evaluation of machine translation systems : sentence level*. In Proceedings of the MT Evaluation

Workshop at the Machine Translation Summit VIII, 22 September, Santiago de Compostela, Spain, pp. 39-43.

MARTIN, Laura (1986). "*Eskimo words for snow": A case study in the genesis and decay of an anthropological example*. American Anthropologist, 88(2), June, pp. 418-423.

MARTINET, André (1960). *Eléments de linguistique générale*. Paris : Armand Colin (1970).

MARTINET, André (1974). *Homonymes et polysèmes*. La linguistique, 10(2), pp. 37-45.

MATES, Benson (1952). *Synonymity*. In LINSKY, Leonard (ed.), pp. 111-136.

MAURANEN, Anna (2002). *Will 'translationese' ruin a contrastive study?* Languages in Contrast, Vol. 2(2), John Benjamins Publishing Company, pp. 161-185.

MAURANEN, Anna & KUJAMAKI, Pekka (eds.) (2004). *Translation Universals: Do They Exist?*, John Benjamins Publishing Company, Philadelphia, PA, USA, 2004.

McCARTHY, Diana, KOELING, Rob, WEEDS, Julie & CARROLL, John (2004a). *Finding Predominant Word Senses in Untagged Text*. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04), 21-26 July, Barcelona, Spain, pp. 279-286.

McCARTHY, Diana, KOELING, Rob, WEEDS, Julie & CARROLL, John (2004b). *Using Automatically Acquired Predominant Senses for Word Sense Disambiguation*. In Proceedings of the ACL SENSEVAL-3 workshop, Barcelone, Espagne, pp. 151-154.

McQUEEN (1967). *Some methods for classification and analysis of multivariate observations*. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

MEDIN, Douglas L. & GOLDSTONE, Robert L. (1995). *The predicates of similarity*. In CACCIARI, Cristina (ed.), pp. 83-110.

MELAMED, Dan I. (1996). *A Geometric Approach to Mapping Bitext Correspondence*. In Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), 17-18 May, Philadelphia, PA, pp. 1-12.

MELAMED, Dan I. (1997). *A Word-to-Word Model of Translational Equivalence*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), Madrid, Spain, pp. 490-497.

MELAMED, Dan I. (1998). *Manual Annotation of Translational Equivalence: The Blinker Project*. Technical Report 98-07, Institute for Research in Cognitive Science, Philadelphia.

MELAMED, Dan I. (2000). *Models of Translational Equivalence among Words*. Computational Linguistics, 26(2), June, pp. 221-249.

MERKEL, Magnus (1988). *Consistency and Variation in Technical Translation, A Study of Translators' Attitudes*. In BOWKER, Lynne, CRONIN, Michael, KENNY, Dorothy, PEARSON, Jennifer (eds.), pp. 137-149.

MEYER, David E., SCHVANEVELDT, Roger W. & RUDDY, Margaret G. (1975). *Loci of Contextual Effects on Visual Word-Recognition*. In RABBITT, Patrick M.A. & DORNIC, Stanislav (eds.), pp. 98-118.

MIHALCEA, Rada & MOLDOVAN, Dan I. (1999). *A Method for Word Sense Disambiguation of Unrestricted Text*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, June, pp. 152-158.

MIHALCEA, Rada & MOLDOVAN, Dan I. (2001). *Automatic generation of a coarse grained WordNet*. In Proceedings of the 14th International Florida Artificial Intelligence Research Society Conference (FLAIRS), 21-23 May, pp. 454-458.

MIHALCEA, Rada, COURTNEY, Corley & STRAPPARAVA, Carlo (2006). *Corpus-based and Knowledge-based Measures of Text Semantic Similarity*. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI) Boston, MA June.

MIHALCEA, Rada, CHKLOVSKI, Timothy & KILGARRIFF, Adam (2004). *The Senseval-3 English Lexical Sample Task*. In Proceedings of ACL/SIGLEX Senseval-3, 25-26 July, Barcelona, Spain, pp. 25-28.

MIHALCEA, Rada (2007). *Knowledge-Based Methods for WSD*, In AGIRRE, Eneko et EDMONDS, Philip (eds.), pp. 107-131.

MIHALTZ, Marton (2005). *Towards a Hybrid Approach To Word-Sense Disambiguation In Machine Translation*. In Proceedings of Workshop on *Modern Approaches in Translation Technologies*, Borovets, Bulgaria.

MILLER, George A., BECKWITH, Richard, FELLBAUM, Christiane, GROOS, Derek & MILLER, Katherine (1990). *Introduction to WordNet: An On-line Lexical Database*. International Journal of Lexicography 3(4), pp. 235-312.

MILLER, George A. & CHARLES, Walter G. (1991). *Contextual correlates of semantic similarity*. Language and Cognitive Processes, 6(1), pp. 1-28.

MILLER, George A., LEACOCK, Claudia, RANDEE Tengi & BUNKER, Ross T. (1993). *A semantic concordance*. In Proceedings of the 3rd DARPA Workshop on Human Language Technology, Princeton, New Jersey, pp. 303-308.

MINSKY, Marvin M. (1968). *Semantic Information Processing*. Cambridge, MA: MIT Press.

MITCHELL, Tom M. (1997). *Machine Learning,* McGraw-Hill International Editions.

MOONEY, Raymond, J. (1996). *Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'96), 17-18 May, Philadelphia, PA, pp. 82-91.

MOUNIN, Georges (1963). *Les problèmes théoriques de la traduction*. Paris : Editions Gallimard.

NASKAR, Sudip Kumar & BANDYOPADHYAY, Sivaji (2007). *JU-SKNSB: Extended WordNet Based WSD on the English All-Words Task at SemEval-1*. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 203-206.

NAVIGLI, Roberto (2006). *Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation* Performance. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING/ACL'06), Sydney, Australia, pp. 105-112.

NAVIGLI, Roberto, LITKOWSKI, Kenneth C. & HARGRAVES, Orin (2007). *SemEval-2007 Task 07 : Coarse-Grained English All-Words Task*. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 30-35.

NEUBERT, Albrecht & SHREVE, Gregory (1992). Translation as Text. Kent (Ohio); London: Kent State University Press

NG, Hwee Tou, CHUNG, Yong Lim & SHOU, King Foo (1999). *A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation*. In Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99), College Park, Maryland, pp. 9-13.

NG, Hwee Tou, WANG, Bin & CHAN, Yee Seng (2003). *Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan, pp. 455-462.

NG, Hwee Tou & CHAN, Yee Seng (2007). *SemEval-2007 Task 11: English Lexical Sample Task via English-Chinese Parallel Text*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 54-58.

NIEßEN, Sojia & HERMANN, Ney (2000). *Improving SMT quality with morpho-syntactic analysis*. In Proceedings of the 18th International Conference on Computational Linguistics, (COLING'00), Saarbrücken, Germany, pp. 1081-1085.

NIEßEN, Sojia & HERMANN, Ney (2004). *Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information*. Computational Linguistics, 30(2), pp. 181-204.

NIRENBURG, Sergei, SOMERS, Harold L. & WILKS, Yorick (eds.) (2003). *Readings in Machine Translation*, MIT Press.

NIWA, Yoshiki & NITTA, Yoshihiko (1994). *Coocurrence vectors from corpora vs distance vectors from dictionaries*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), 5-9 August, Kyoto, Japan, pp. 304-309.

NUNBERG, Geoffrey (1979). *The non-uniqueness of semantic solutions: polysemy*. Linguistics and Philosophy 3(2), Dordrecht, Holland / Boston, U.S.A.: D. Reidel Publishing Company, pp. 143-184.

NUNBERG, Geoffrey & ZAENEN, Annie (1992). *Systematic Polysemy in Lexicology and Lexicography*. In Proceedings of EURALEX'92, University of Tampere, Finland, pp. 387-395.

OCH, Franz Josef & WEBER, Hans (1998). *Improving Statistical Natural Language Translation with Categories and Rules*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL'98), Montreal, Quebec, Canada, 10-14 August, pp. 985-989.

OCH, Franz Josef (1999). *An Efficient Method for Determining Bilingual Word Classes*. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99), 8-12 June, Bergen, Norway, pp. 71-76.

OCH, Franz Josef, TILLMANN, Christoph & HERMANN, Ney (1999). *Improved Alignment Models for Statistical Machine Translation*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 21-22 June, University of Maryland, College Park, pp. 20-28.

OCH, Franz Josef & HERMANN, Ney (2000). *Improved Statistical Alignment Models*, In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00), 3-6 October, Hong Kong, China, pp. 440-447.

OCH, Franz Josef & HERMANN, Ney (2003). *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, 29(1), pp. 19-51.

OCH, Franz Josef et HERMANN, Ney (2004). *The alignment template approach to statistical machine translation*. Computational Linguistics, 30(4), pp. 417-449.

OLOHAN, Maeve (2004). *Introducing corpora in Translation Studies*. London: Routledge.

OSWALD, Victor A. (1952). *Microsemantics*. Paper presented at the first MT conference, June, MIT (not published).

OZDOWSKA, Sylwia (2006). *Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora*. In Proceedings of the EACL Workshop on Cross-Language Knowledge Induction, 3 April, Trento, Italy, pp. 53-60.

PALMER, Martha, NG Hwee Tou & DANG, Hoa Trang (2007). *Evaluation of WSD Systems*. In AGIRRE, Eneko et EDMONDS, Philip (eds.), pp. 75-106.

PANTEL, Patrick & LIN, Dekang (2002). *Discovering word senses from text*. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 23-26 June, Edmonton, Alberta, Canada, pp. 613-619.

PANTEL, Patrick & LIN, Dekang (2003). *Automatically Discovering Word Senses.* In Proceedings of Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL), Demonstrations, 27 May-1 June, Edmonton, Alberta, Canada, pp. 21-22.

PAPAGEORGIOU, Harris, CRANIAS, Lambros & PIPERIDIS, Stelios (1994). *Automatic alignment in parallel corpora*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), 27-30 June, Las Cruces, New Mexico State University, pp. 334-336.

PAPAGEORGIOU, Harris, PROKOPIDIS, Prokopis, GIOULI, Voula & PIPERIDIS, Stelios (2000). *A Unified POS Tagging Architecture and its Application to Greek*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), 31 May-2 June, Athens, Greece, pp. 1455-1462.

PAPINENI, Kishore, ROUKOS Salim, WARD, Todd & ZHU, Wei-Jing (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of the

40th Annual Meeting of the Association for Computational Linguistics (ACL'02), 6-12 July, Philadelphia, pp. 311-318.

PAPROTTÉ, Wolf (1998). *Word Sense Disambiguation: An Experimental Study for Germans*. In WEIGAND, Edda (ed.), pp. 243-261.

PATWARDHAN, Siddharth (2003). *Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness*. Master's Thesis, University of Minnesota.

PATWARDHAN, Siddharth, SATANJEEV, Banerjee & PEDERSEN, Ted (2003). *Using measures of semantic relatedness for word sense disambiguation*. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Mexico City, February 16-22, pp. 241-257.

PATWARDHAN, Siddharth (2006). *Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts*. In Proceedings of EACL 2006, Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy, April 4, pp. 1-8.

PEDERSEN, Ted & BRUCE, Rebecca (1997a). *Distinguishing Word Senses in Untagged Text*. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP'97), August 1-2, Providence, R.I, pp. 197-207.

PEDERSEN, Ted & BRUCE, Rebecca (1997b). *A new supervised learning algorithm for word sense disambiguation*. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, Providence, RI, July 27-31, pp. 604-609.

PEDERSEN, Ted, BRUCE, Rebecca & WIEBE, Janyce (1997). *Sequential model selection for word sense disambiguation*. In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, March 31 – April 3, pp. 388-395.

PEDERSEN TED & BRUCE, Rebecca (1998). *Knowledge lean word sense disambiguation*. In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI'98), Madison, Wisconsin, July 26-30, pp. 800-805.

PEDERSEN, Ted, PATWARDHAN, Siddharth & MICHELIZZI, Jason (2004). *WordNet::Similarity – Measuring the Relatedness of Concepts*. In Proceedings of the Nineteenth National Conference on Artificial Intelligence (Intelligent Systems Demonstrations) (AAAI'04), San Jose, California, July 25-29, pp. 1024-1025.

PEDERSEN, Ted (2007). *Unsupervised Corpus-Based Methods for WSD*. In AGIRRE, Eneko & EDMONDS, Philip (eds.), pp. 133-166.

PEREIRA, Fernando & TISHBY, Naftali (1992). *Distributional similarity, phase transitions and hierarchical clustering*. Working Notes of the AAAI Symposium on Probabilistic Approaches to Natural Language, Cambridge, MA, October 23-25, pp. 108-112.

PEREIRA, Fernando, TISHBY, Naftali & LEE, Lillian (1993). *Distributional Clustering of English Words*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), Ohio State University, Columbus, Ohio, June 22-26, pp. 183-190.

PERFETTI, Charles A. (1998). *The Limits of Co-Occurrence: Tools and Theories in Language Research*. Discourse Processes, 25(2&3), pp. 363-377.

PETERS, Wim, PETERS, Ivonne & VOSSEN, Piek (1998). *Automatic sense clustering in EuroWordNet*. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain, May 28-30, pp. 409-416.

PICHON, Ronan & SEBILLOT, Pascale (1999). *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience*. In Proceedings of the 6e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN'99), Cargèse, France.

PIPERIDIS, Stelios, MALAVAZOS, Christos & TRIANTAFYLLOU, Ioannis (1999). *A Multi-level Framework for Memory-Based Translation Aid Tools*. Aslib, Translating and the Computer 21, London.

PIPERIDIS, Stelios, DIMITRAKIS, Panagiotis & BALTA, Irene (2005). *Lexical Transfer Selection Using Annotated Parallel Corpora*. In Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing (RANLP), September 21-23, Borovets, Bulgaria.

PLOUX, Sabine & JI, Hyungsuk (2003). *A Model for Matching Semantic Maps between Languages (French/English, English/French)*. Computational Linguistics 29(2), pp. 155-178.

POLGUERE, Alain (2003). *Lexicologie et sémantique lexicale: notions fondamentales*, coll. « Paramètres », Montréal: Presses de l'Université de Montréal.

PRADHAN, Sameer S., LOPER, Edward, DLIGACH, Dmitriy & PALMER, Martha (2007). *Sem-Eval-2007 Task 17: English Lexical Sample, SRL and All Words*. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, June 23-24, pp. 87-92.

PULLUM, Geoffrey K. (1991). *The great Eskimo vocabulary hoax*. In *The great Eskimo vocabulary hoax and other irreverent essays on the study of language*. Chicago: University of Chicago Press, pp. 159-171.

PURANDARE, Amruta (2004). *Word Sense Discrimination by Clustering Similar Contexts*. Master of Science Thesis, Department of Computer Science, University of Minnesota, Duluth, August.

PURANDARE, Amruta & PEDERSEN, Ted (2004a). *Discriminating Among Word Meanings By Identifying Similar Contexts*. In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), San Jose, California, July 25-29, pp. 964-965.

PURANDARE, Amruta & PEDERSEN, Ted (2004b). *Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces*. In Proceedings of the Conference on Computational Natural Language Learning (CONLL'04), Boston, MA, May 6-7, pp. 41-48.

PUSTEJOVSKY, James (1995). *The Generative Lexicon*, MIT Press.

PUSTEJOVSKY, James & BOGURAEV, Branimir (1996a). *Introduction: Lexical Semantics in Context*. In PUSTEJOVSKY, James & BOGURAEV, Branimir (ed.), pp. 1-14.

PUSTEJOVSKY, James & BOGURAEV, Branimir (ed.) (1996b). *Lexical Semantics, The Problem of Polysemy*, Clarendon Paperbacks, Oxford.

QUILLIAN, M. Ross (1968). *Semantic memory*. In MINSKY, Marvin (ed.), pp. 227-270.

QUINE, Willard Van Orman (1953). *The problem of Meaning in Linguistics*. In *From a Logical Point of View: nine Logico-Philosophical Essays*, Harvard University Press, Cambridge, Massachusetts, pp. 47-64.

RABBITT, Patrick M.A. & DORNIC, Stanislav (eds.) (1975). *Attention and performance V*. London: Academic Press.

RADA, Roy, MILI, Hafedh, BICKNELL, Ellen & BLETTNER, Maria (1989). *Development and application of a metric on semantic nets*. IEEE Transaction on Systems, Man, and Cybernetics, 19(1), pp. 17-30.

RAPP, Reinhard (1995). *Identifying word translations in non-parallel texts*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95), Cambridge, MA, June 26-30, pp. 320-322.

RAPP, Reinhard (2003). *Word Sense Discovery Based on Sense Descriptor Dissimilarity*. In Proceedings of the 9th Machine Translation Summit (MT Summit), New Orleans, September 23-27, pp. 315-322.

RASTIER, François, CAVAZZA, Marc & ABEILLE, Anne (1994). *Sémantique pour l'analyse : de la Linguistique à l'Informatique*. Masson, Paris.

REIFLER E. (1954). *The first conference on mechanical translation*. Mechanical Translation 1(2), pp. 23-32.

REITER, Ehud & SRIPADA, Somayajulu (2004). *Contextual Influences on Near-Synonym Choice*. In Proceedings of the Third International Conference on Natural Language Generation (INLG'04), University of Brighton, July 14-16, pp. 161-170.

RESNIK, Philip (1995). *Using information content to evaluate semantic similarity in a taxonomy*. In Proceedings of the International Joint conference for Artificial Intelligence (IJCAI-95), Montreal, Canada, August 20-25, pp. 448-453.

RESNIK, Philip (1999). *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*. Journal of Artificial Intelligence Research (JAIR), 11, pp. 95-130.

RESNIK, Philip (2004). *Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation*. In GELBUKG, Alexander (ed.), Lecture Notes in Computer Science (2945): Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Proceedings, Seoul, Korea, February 15-21, Springer, pp. 283-299.

RESNIK, Philip (2007). *WSD in NLP Applications*. In AGIRRE, Eneko & EDMONDS, Philip (eds.), pp.299-337.

RESNIK, Philip & YAROWSKY, David (1997). *A Perspective on Word Sense Disambiguation Methods and Their Evaluation*. In Proceedings of SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?", Washington, D.C., April 4-5, pp. 79-86.

RESNIK, Philip & YAROWSKY, David (2000). *Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation*. Natural Language Engineering 5(3), Cambridge University Press, pp. 113-133.

RESNIK, Philip & DIAB, Mona (2002). *An Unsupervised Method for Word Sense Tagging using Parallel Corpora*. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA, July 6-12, pp. 255-262.

RICHARDSON R., SMEATON A.F. & J. MURPHY (1995). *Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words*. Working Paper: CA-0395, School of Computer Applications, Dublin City University.

ROSSIGNOL, Mathias & SEBILLOT, Pascale (2006). *Mise au jour semi-automatique de nuances sémantiques entre mots de sens proches*. In Proceedings of TALN'06, Leuven, Belgium, April 10-13, pp. 266-275.

RUBENSTEIN, Herbert & GOODENOUGH, John B. (1965). *Contextual correlates of synonymy*. Communications of the ACM, 8(10), pp. 627-633.

SALKIE, Raphael (1995). *Parallel Corpora, Translation Equivalence and Contrastive Linguistics*. Paper read at the Association for Literary and Linguistic Computing / Association for Computing in the Humanities Joint International Conference, University of California, Santa Barbara, July 11-15.

SALKIE, Raphael (1997). *Naturalness and contrastive linguistics,* dans LEWANDOWSKA-TOMASZCZYK, Barbara & MELIA, James Patrick (eds.). In Proceedings of PALC'97: Practical Applications in Language Corpora, Lodz: Lodz University Press, April 10-14, pp. 297-312.

SALKIE, Raphael (2002a). *How can linguists profit from parallel corpora?* In Lars Borin (ed.), pp. 93-109.

SALKIE, Raphael (2002b). *Quelques questions méthodologiques dans l'exploitation des corpus multilingues*. In BILGER, Mireille (ed.), *Corpus, méthodologie et applications linguistiques*, pp. 180-195.

SALKIE, Raphael (2002c). *Two types of translation equivalence*. In ALTENBERG, Bengt & GRANGER, Sylviane (eds.), pp. 51-71.

SALTON, Gerard & BUCKLEY, Chris (1990). *Improving retrieval performance by relevance feedback*. Journal of the American Society for Information Science, 41(4), pp. 288-297.

SANTORINI, Beatrice (1991). Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania.

SCHMID, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, Manchester, pp. 44-49.

SCHMIED, Josef & SCHAFFLER, Hildegard (1996). *Approaching translationese through parallel and translational corpora*. In Synchronic corpus linguistics, Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16), PERCY, Carol E., MEYER, Charles F. & LACASHIER, Ian (eds.), Amsterdam : Rodopi, pp. 41-56.

SCHÜTZE, Hinrich (1992). *Dimensions of Meaning*. In Proceedings of Supercomputing'92, Minneapolis, pp. 787-796.

SCHÜTZE, Hinrich (1998). *Automatic Word Sense Discrimination*. Computational Linguistics, Vol. 24, Number 1. pp. 97-123.

SCHÜTZE, Hinrich & PEDERSEN, Jan O. (1995). *Information retrieval based on word senses*. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 24-26 April, pp. 161-175.

SEGOND, Frédérique (2000). *Framework and Results for French*. Computers and the Humanities, 34, pp. 49-60.

SEGOND, Frédérique, AIMELET, Elisabeth, LUX, Veronika & JEAN, Corinne (2000). *Dictionary-Driven Semantic Look-up*. Computers and the Humanities, Special Issue on Senseval, 34, pp. 193-197.

SEIDENBERG M.S., WATERS G.S., SANDERS M. & LANGER PP. (1984). *Pre- and postlexical loci of contextual effects on word recognition*. Memory & Cognition, vol. 12, no 4, pp. 315-328.

SHAPIRO, Stuart Charles (1992). *Encyclopedia of artificial intelligence*, New York: John Wiley & Sons, Inc.

SHEPARD, Roger N. (1962). *The analysis of proximities: Multidimensional scaling with an unknown distance function*. Psychometrika, 27(2), pp. 125-140.

SHLESINGER, Miriam (1991). *Lexicalization in translation: an empirical study of students' progress*. In DOLLERUP, Cay & LODDEGAARD, Anne (eds.), pp. 123-127.

SIMARD, Michel & LANGLAIS, Philippe (2003). *Statistical Translation Alignment with Compositionality Constraints*. In Proceedings of the HLT-NAACL Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Canada, May 31, pp. 19-22.

SIMARD, Michel, FOSTER, George F. & ISABELLE, Pierre (1992). *Using Cognates to Align Sentences in Bilingual Corpora*. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Montreal, Canada, June 25-27, pp. 67-81.

SIMPSON, Greg B. (1989). *Varieties of Ambiguity: What Are We Seeking?* In GORFEIN, David S. (ed.), *Resolving Semantic Ambiguity*, pp. 13-21.

SINCLAIR, John M. (1996). An *International Project in Multilingual Lexicography*. International Journal of Lexicography, 9(3), pp. 179-196.

SINCLAIR, John M., PAYNE, Jonathan & PEREZ HERNANDEZ, Chantal (1996). *Introduction*. In SINCLAIR, John M., PAYNE, Jonathan & PEREZ HERNANDEZ, Chantal (eds.), *Corpus to Corpus: A Study of Translation Equivalence*. Special Issue. International Journal of Lexicography, 9(3), Oxford University Press, September, pp. i-ix.

SINHA, Ravi & MIHALCEA, Rada (2007). *Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity*. In Proceedings of the First IEEE International Conference on Semantic Computing (ICSC'07), Irvine, California, September 17-19, pp. 363-369.

SLAKTA, Denis (1985). *Grammaire de texte : synonymie et paraphrase*. In FUCHS, Catherine (ed.), pp. 123-140.

SLATOR, Brian M. (1992). *Sense and Preference*. Computer and Mathematics with Applications, 23(6/9), pp. 391-402.

SMALL, Steven L. (1979). *Word expert parsing*. In Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics (ACL'79), San Diego, La Jolla, California, June 29 – July 1, pp. 9-13.

SNELL-HORNBY, Mary (1990). *Dynamics in meaning as a problem for bilingual lexicography*. In TOMASZCZYK, Jerzy & LEWANDOWSKA-TOMASZCZYK, Barbara (eds.), pp. 209-226.

SNYDER, Benjamin & PALMER, Martha (2004). *The English All-Words Task*. In Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, Spain, July 25-26, pp. 41-43.

SOMERS, Harold (ed.) (1996). *Terminology, LSP and Translation: studies in language engineering in honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

SOVRAN, Tamar (1992). *Between similarity and sameness*. Journal of Pragmatics 18, pp. 329-344.

SPARCK JONES, Karen (1964). *Synonymy and Semantic Classification*. Ph.D. Thesis, University of Cambridge. Reprinted in 1986 by Edinburgh University Press, Edinburgh.

SPECIA, Lucia (2005). *A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation*. In Proceedings of the 8th Research Colloquium of the UK Special-interest Group in Computational Linguistics (CLUK-05), Manchester, January 11, pp. 71-78.

SPECIA, Lucia, DAS GRAÇAS VOLPE NUNES, Maria & STEVENSON, Marc (2005). *Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation*. In Proceedings of Recent Advances in Natural Language Processing (RANLP'05), September 21-23, Borovets, Bulgaria.

SPECIA, Lucia, DAS GRAÇAS VOLPE NUNES, Maria & STEVENSON, Marc (2006a). *Translation Context Sensitive WSD*. In Proceedings of the 11th Annual

Conference of the European Association for Machine Translation (EAMT'06), Oslo, Norway, June 19-20, pp. 227-232.

SPECIA, Lucia, DAS GRAÇAS VOLPE NUNES, Maria, CASTELO BRANCO RIBEIRO, Gabriela & STEVENSON, Marc (2006b). *The Need for Application-Dependent WSD Strategies: a Case Study in MT*. In Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese (Propor'06), LNAI 3960, May 13-17, Itatiaia, pp. 233-237.

SPECIA, Lucia, DAS GRAÇAS VOLPE NUNES, Maria, RIBEIRO GABRIELA CASTELO BRANCO & STEVENSON, Marc (2006c). *Multilingual versus Monolingual WSD*. In Proceedings of the Workshop Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together – EACL-2006, April 3-7, Trento, Italy, pp. 33-40.

SPECIA, Lucia, DAS GRAÇAS VOLPE NUNES, Maria & STEVENSON, Marc (2007). *Learning Expressive Models for Word Sense Disambiguation*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic, 23-30 June, pp. 41-48.

SPENCE, Donald P. & OWENS, Kimberly C. (1990). *Lexical co-occurrence and association strength*. Journal of Psycholinguistic Research, 19(5), pp. 317-330.

SPERBER, Dan & WILSON, Deirdre (1989). *La pertinence, Communication et cognition*. Paris : Les éditions de minuit.

STAMOU, Sofia, OFLAZER Kemal, PALA Karel, CHRISTODOULAKIS, Dimitris, CRISTEA Dan, TUFIS, Dan, KOEVA, Svetla, TOTKOV, George, DUTOIT, Dominique & GRIGORIADOU, Maria (2002). *BALKANET : A Multilingual Semantic Network for the Balkan Languages*. In Proceedings of the International Wordnet Conference, Janyary 21-25, Mysore, India, pp. 12-14.

STAMOU, Sofia, NENADIC Goran & CHRISTODOULAKIS, Dimitris (2004). *Exploring Balkanet Shared Ontology for Multilingual Conceptual Indexing*. In Proceedings of the 4th Language Resources and Evaluation Conference (LREC), May 26-28, Lisbonne, Portugal, pp. 781-784.

STEDE, Manfred (1999). *Lexical Semantics and Knowledge Representation in Multilingual Text Generation*. Boston/Dordrecht/London: Kluwer Academic Publishers.

SUSSNA, Michael (1993). *Word Sense Disambiguation for free-text indexing using a massive semantic network*. In Proceedings of the Second International Conference on Information and Knowledge Base Management (CIKM'93), Arlington, Virginia, pp. 67-74.

TARSKI, Alfred (1944). *The Semantic Conception of Truth and the Foundations of Semantics*. Philosophy and Phenomenological Research, 4, pp. 341-376.

TABOSSI, Patrizia (1989). *What's in a Context?* In GORFEIN David S. (ed.), *Resolving Semantic Ambiguity*, pp. 25-39.

TEICH, Elke (2002). *System-oriented and text-oriented comparative linguistics research*. Languages in Contrast, Vol. 2, Number 2, John Benjamins Publishing Company, pp. 187-210.

TERRA, Egidio (2004). *Lexical Affinities and Language Applications*. PhD Thesis, University of Waterloo, Ontario, Canada.

TEUBERT, Wolfgang (1996). *Comparable or Parallel Corpora?* International Journal of Lexicography, Vol. 9 (3), pp. 238-264.

TEUBERT, Wolfgang (2002). *The role of parallel corpora in translation and multilingual lexicography*. In ALTENBERG, Bengt & GRANGER, Sylviane (eds.), *Lexis in contrast: Corpus-based approaches*, pp. 189-214.

THOMPSON, Henry S. (1991). *Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment*. In Proceedings of the Evaluators' Forum, Kirsten Falkedal (ed.), Geneva: ISSCO, April 21-24, Les Rasses, Vaud, Suisse, pp. 215-223.

THOUARD, Denis (ed.) (2000). WILHELM VON HUMBOLDT, *Sur le caractère national des langues et autres écrits sur le langage*. Paris : Editions du Seuil.

THUE VOLD, Eva, LYSE, Gunn Inger & MULLER GJESDAL, Anje (2006). (eds.) *New Voices in Linguistics*. Cambridge Scholars Publishing, Newcastle.

THUNES, Martha (2003). *Evaluating thesaurus entries derived from translational features*. In Proceedings of the 14th Nordic Conference on Computational Linguistics (Nodalida), Reykjavík, May 30-31.

THUNES MARTHA (2004). *Extracting lexical translation correspondents from parallel text*, draft.

TIEDEMANN, Jörg (2001). *Predicting Translations in Context*. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'01), Tzigov Chark, Bulgaria, 5-7 September, pp. 240-244.

TIEDEMANN, Jörg (2003). *Combining clues for word alignment*. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary, April 12-17, pp. 339-346.

TIEDEMANN, Jörg (2004). *Word to word alignment strategies*. In Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), Geneva, Switzerland, August 23-27, pp. 212-218.

TIRKKONEN-CONDIT, Sonja (2002). *Translationese – a myth or an empirical fact?* Target 14(2), pp. 207-220.

TIRKKONEN-CONDIT, Sonja (2004). *Unique items – over- or under- represented in translated language?*,In MAURANEN, Anna & KUJAMAKI, Pekka (eds.), pp. 177-184.

TOGNINI-BONELLI, Elena (1996). *Towards Translation Equivalence from a Corpus Linguistics Perspective*. International Journal of Lexicography, Vol. 9, No. 3.

TOMASZCZYK, Jerzy & LEWANDOWSKA-TOMASZCZYK, Barbara (1990) (eds.), *Meaning and Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Co.

TRIANTAFYLLOU, Ioannis, DEMIROS, Iason, MALAVAZOS, Christos & PIPERIDIS, Stelios (2000). *An alignment architecture for Translation Memory Bootstrapping*. In Proceedings of MT2000, Exeter, UK.

TUFIS, Dan, CRISTEA, Dan & STAMOU, Sofia (2004a). *BalkaNet : Aims, Methods, Results and Perspectives, A General Overview*. Romanian Journal of Information Science and Technology, TUFIS, Dan (ed.), Special Issue on Balkanet, Romanian Academy, 7(1-2), pp. 9-43.

TUFIS, Dan, ION, Radu & IDE, Nancy (2004b). *Word Sense Disambiguation as a Wordnet's Validation Method in Balkanet*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, May 26-28, pp. 1071-1074.

TUFIS, Dan, ION, Radu & IDE, Nancy (2004c). *Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets*. In Proceedings of the 20th International Conference on Computational Linguistics, (COLING'04), Geneva, Switzerland, pp. 1312-1318.

TUGGY, David (1993). *Ambiguity, polysemy, and vagueness*. Cognitive Linguistics 4(3), pp. 273-290.

TURIAN, Joseph P., SHEN, Luke & MELAMED, Dan I. (2003). *Evaluation of Machine Translation and its Evaluation*. In Proceedings of the 9th Machine Translation Summit, New Orleans, September 23-27, pp. 23-38.

TURNEY, Peter (2001). *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*. In Proceedings of the Twelfth European Conference on Machine Learning

(ECML-2001), DE RAEDT, Luc & FLACH, Peter (eds.), Freiburg, Germany, pp. 491-502.

TVERSKY, Amos (1977). *Features of similarity*. Psychological Review, 84, pp. 327-352.

TYMOCZKO, Maria (2004). *Difference in Similarity*. In Proceedings of the International Conference on Similarity and Translation, New York, May 31 – June 1 2001, ARDUINI, Stefano & HODGSON Robert (eds.), Guaraldi, Rimini, Italy.

ULLMANN, Stephen (1962). *Semantics, An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.

VAN DER PLAS, Lonneke & BOUMA, Gosse (2004). *Syntactic Contexts for Finding Semantically Related Words.* In Proceedings of the Meeting of Computational Linguistics in the Netherlands (CLIN'04), Leiden University, December 17, pp. 173-186.

VAN DER PLAS, Lonneke & TIEDEMANN, Jörg (2006). *Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity.* In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING'06), Sydney, Australia, July 17-21, pp. 866-873.

VAN RIJSBERGEN Cornelis Joost (1979). *Information retrieval*. (2e edition). Butterworths, London.

VASILESCU, Florentina, LANGLAIS, Philippe & LAPALME, Guy (2004). *Evaluating Variants of the Lesk Approach for Disambiguating Words*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, May 24-31, pp. 633-636.

VERDEJO, Felisa, GONZALO, Julio, PENAS, Anselmo, LOPEZ, Fernando & DAVID FERNANDEZ, David (2000). *Evaluating wordnets in Cross-Language Information Retrieval: the ITEM search engine*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00), Athens Greece, pp. 1769-1774.

VERONIS, Jean (1998). *A study of polysemy judgements and inter-annotator agreement*. In Programme and advanced papers of the Senseval workshop, Herstmonceux Castle, Angleterre.

VERONIS, Jean & IDE, Nancy (1990). *Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries*. In Proceedings of the 13rd International Conference on Computational Linguistics (COLING'90), Helsinki, August 20-25, Vol. 2, pp. 389-394.

VERONIS, Jean & LANGLAIS, Philippe (1999). *ARCADE : évaluation de systèmes d'alignement de textes multilingues*. In CHIBOUT, Karim, MARIANI, Joseph, MASSON, Nicolas & NEEL, Françoise (eds.), pp. 77-100.

VERONIS, Jean & LANGLAIS, Philippe (2000). *Evaluation of parallel text alignment systems. The ARCADE project*. In Jean Véronis (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer Academic Publishers, pp. 369-388.

VERONIS, Jean (2003). *Hyperlex : cartographie lexicale pour la recherche d'informations*. In Proceedings of TALN'03, Batz-sur-mer, France, pp. 265-274.

VERONIS, Jean (2004). *Hyperlex : lexical cartography for information retrieval*. Computer, Speech and Language, Special Issue on Word Sense Disambiguation, 18(3), pp. 223-252.

VIBERG, Åke (2002). *Polysemy and disambiguation cues across languages: the case of Swedish* få *and English* get. In ALTENBERG, Bength & GRANGER, Sylviane (eds.), pp. 119-150.

VICKREY, David, BIEWALD, Luke, TEYSSIER, Marc & KOLLER, Daphne (2005). *Word-Sense Disambiguation for Machine Translation*. In Proceedings of the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP'05), October 6-8, Vancouver, Canada, pp. 771-778.

VICTORRI, Bernard & FUCHS, Catherine (1996). *La polysémie : construction dynamique du sens*. Paris : Hermès.

VICTORRI, Bernard (1997). *La polysémie: un artefact de la linguistique?* Revue de Sémantique et de Pragmatique, 2, pp. 41-62.

VOGEL, Stephan, NEY, Hermann & TILLMANN, Christoph (1996). *HMM-Based Word Alignment in Statistical Translation*. In Proceedings of the 16th International Conference on Computational Linguisics (COLING'96), Copenhagen, Denmark, August 5-9, pp. 836-841.

VOORHEES, Ellen M. (1993). *Using WordNet to disambiguate word senses for text retrieval*. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrievali, Pittsburgh, PA, June 27 – July 1, pp. 171-180.

VOSSEN, Piek (ed.) (1999). *EuroWordNet General Document*. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document.

VOSSEN, Piek, PETERS, Wim & GONZALO, Julio (1999). *Towards a Universal Index of Meaning*. In Proceedings of ACL-99 Workshop, Siglex-99, Standardizing

Lexical Resources, June 21-22, University of Maryland, College Park, Maryland, pp. 81-90.

WALTZ, David L. & POLLACK, Jordan B. (1985). *Massively parallel parsing: A strongly interactive model of natural language interpretation*. Cognitive Science, 9, pp. 51-74.

WEAVER, Warren (1949). *Translation*. Reprinted in Readings in Machine Translation, Nirenburg *et al*. (eds.) (2003), MIT Press.

WEEDS, Julie (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD Thesis, University of Sussex, September.

WEEDS, Julie & WEIR, David (2005). *Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity*. Computational Linguistics, 31(4), December, pp. 439-475.

WEEDS, Julie, WEIR, David & McCARTHY Diana (2004). *Characterising Measures of Lexical Distributional Similarity*. In Proceedings of the 20th International Conference of Computational Linguistiques (COLING'04), Geneva, Switzerland, pp. 1015-1021.

WEIGAND, Edda (1998a). (ed.). *Contrastive Lexical Semantics*, John Benjamins Publishing Company, Amsterdam.

WEIGAND, Edda (1998b). *Contrastive Lexical Semantics.* In WEIGAND EDDA (ed.), pp. 25-44.

WESTHEIDE, Henning (1998). *Equivalence in Contrastive Semantics, The effect of Cultural Differences*. In WEIGAND, Edda (ed.), pp. 119-137.

WHITE, John S., O'CONNELL, Theresa & O'MARA, Francis (1994). *The ARPA MT evaluation methodologies : evolution, lessons, and future approaches*. In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA), October 5-8, Columbia, Maryland, pp. 193-205.

WHORF, Benjamin Lee (1940). *Science and linguistics*. Technology Review (MIT) 42(6), April, pp. 229-31, pp. 247-68. Reprinted in CAROLL, John B. (ed.), *Language, Thought, and Reality*, pp. 207-219.

WHORF, Benjamin Lee (1969). *Linguistique et Anthropologie*. Denoël / Gonthier, Paris (traduction de *Language, Thought and Reality*, 1956, MIT Press, Cambridge Massachusetts).

WIDDOWS, Dominic & DOROW, Beate (2002). *A Graph Model for Unsupervised Lexical Acquisition*. 19th International Conference on Computational

Linguistics (COLING'02), Taipei, Taiwan, August 24 – September 1, pp. 1093-1099.

WIERZBICKA, Anna (1992). *Semantics, Culture, and Cognition*: *Universal Human Concepts in Culture-Specific Configurations*. Oxford University Press, New York, Oxford.

WILKS, Yorick (1975). *An Intelligent Analyzer and Understander of English*. Communications of the ACM, 18(5), pp. 264-274.

WILKS, Yorick, FASS, Dan, GUO Cheng-Ming, MCDONALD, James E., PLATE, Tony & SLATOR, Brian M. (1990). *Providing machine tractable dictionary tools*. Machine Translation, Vol. 5(2), pp. 99-154.

WILKS, Yorick & FASS, Dan (1992). *Preference semantics: A family history*. In SHAPIRO, Stuart C. (ed.), *The Encyclopedia of Artificial Intelligence*.

WILKS, Yorick & STEVENSON, Mark (1996). *The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging?* Technical Report CS-96-05, University of Sheffield.

WITTEN, Ian H. & FRANK, Eibe (2005). *Data Mining*, *Practical Machine Learning Tools and Techniques*. 2nd edition, Amsterdam/Boston/Paris: Morgan Kaufmann.

WITTGENSTEIN, Ludwig (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.

WOODS, William A. (1997). *Conceptual Indexing: A Better Way to Organize Knowledge*. Technical Report TR-9761, Sun Microsystems Laboratories, Mountain View, CA.

WU, Hua & PALMER, Martha (1994). *Verb semantics and lexical selection*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), Las Cruces, New Mexico, June 27-30, pp. 133-138.

WU, Hua & ZHOU, Ming (2003). *Optimizing Synonym Extraction Using Monolingual and Bilingual Resources*. In Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003), Sapporo, Japan, July, pp. 72-79.

YAROWSKY, David (1992). *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora*. In Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), Nantes, France, August 23-28, pp. 454-460.

YAROWSKY, David (1993). *One sense per collocation*. Proceedings of ARPA Human Language Technology Workshop, Princeton, New Jersey, March 21-23, pp. 266-271.

YAROWSKY, David (1994). *Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), Las Cruces, New Mexico, June 27-30, pp. 88-95.

YAROWSKY, David (1995). *Unsupervised word sense disambiguation rivalling supervised methods*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95), Cambridge, MA, June 26-30, pp. 189-196.

YAROWSKY, David, NGAI, Grace & WICENTOWSKI, Richard (1991). *Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora*. In Proceedings of the 1st International Conference on Human Language Technology Research (HLT 2001), San Diego, California, March 18-21, pp. 161-168.

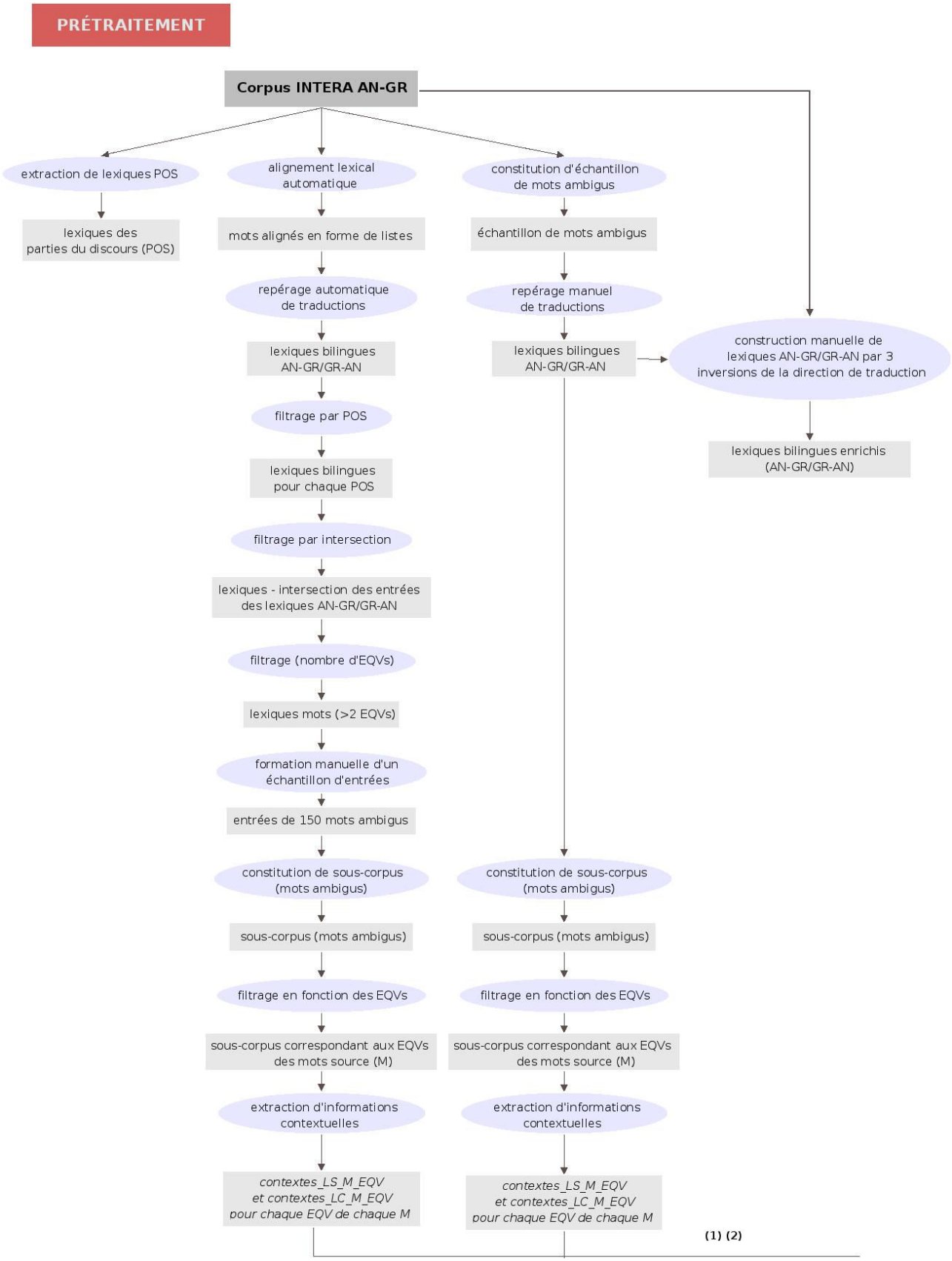
ZIPF, George Kingsley (1945). *The meaning-frequency relationship of words*. Journal of General Psychology, 33, pp. 251-266.

ZIPF, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

ANNEXES

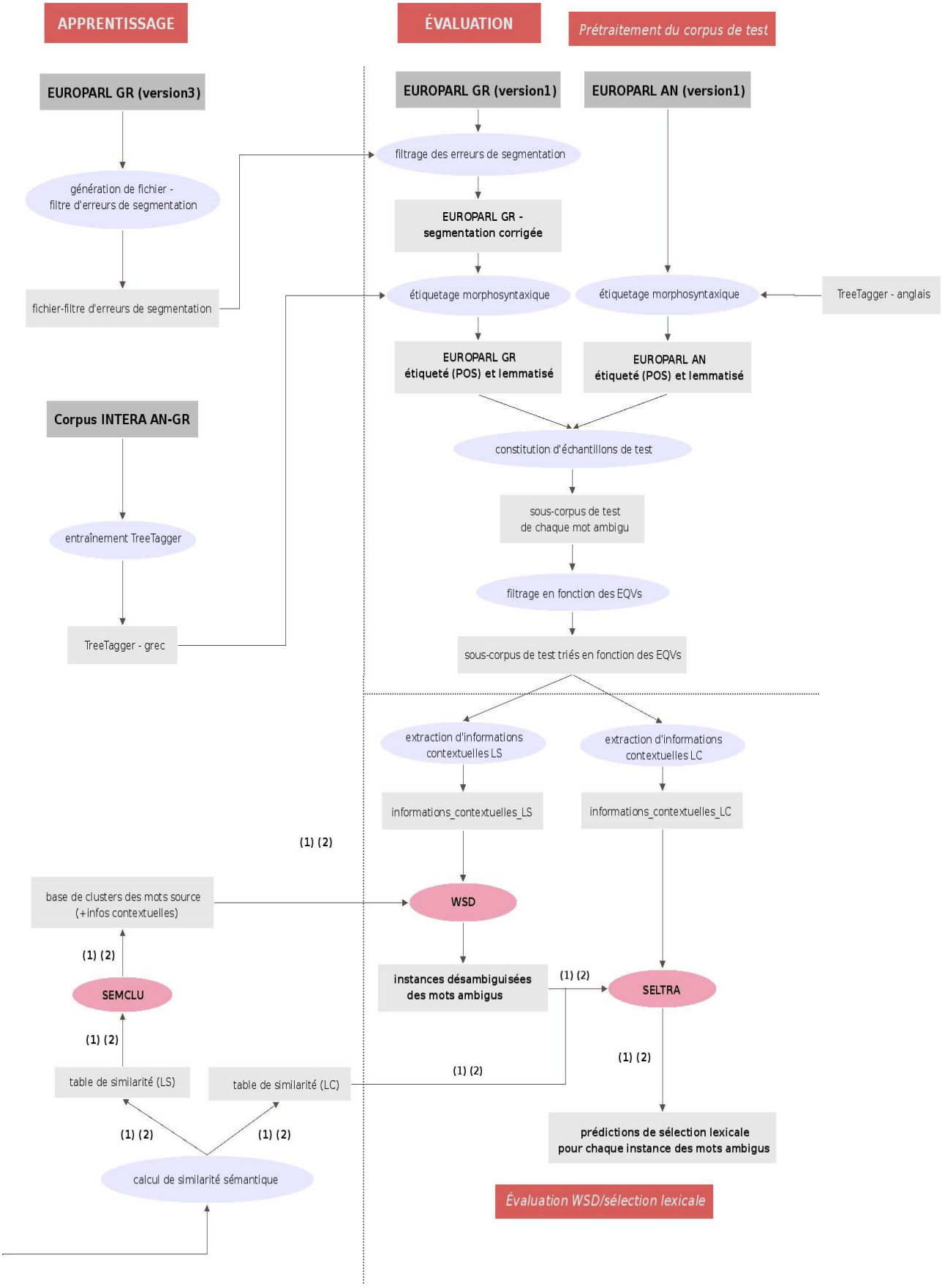
Annexe A

A1. Diagramme de flux de données



(1) : 1re étape, exploitation des lexiques bilingues générés automatiquement
(2) : 2e étape, exploitation des lexiques bilingues générés manuellement

SEMCLU : programme de clustérisation sémantique
WSD : programme de désambiguisation lexicale
SELTRA : programme de sélection lexicale


APPRENTISSAGE
ÉVALUATION
Prétraitement du corpus de test
EUROPARL GR (version3)
EUROPARL GR (version1)
EUROPARL AN (version1)
génération de fichier - filtre d'erreurs de segmentation
filtrage des erreurs de segmentation
EUROPARL GR - segmentation corrigée
fichier-filtre d'erreurs de segmentation
étiquetage morphosyntaxique
étiquetage morphosyntaxique
TreeTagger - anglais
EUROPARL GR étiqueté (POS) et lemmatisé
EUROPARL AN étiqueté (POS) et lemmatisé
Corpus INTERA AN-GR
constitution d'échantillons de test
entraînement TreeTagger
sous-corpus de test de chaque mot ambigu
filtrage en fonction des EQVs
TreeTagger - grec
sous-corpus de test triés en fonction des EQVs
extraction d'informations contextuelles LS
extraction d'informations contextuelles LC
informations_contextuelles_LS
informations_contextuelles_LC
(1) (2)
base de clusters des mots source (+infos contextuelles)
WSD
(1) (2)
SEMCLU
instances désambiguisées des mots ambigus
(1) (2)
SELTRA
(1) (2)
table de similarité (LS)
table de similarité (LC)
(1) (2)
(1) (2)
(1) (2)
prédictions de sélection lexicale pour chaque instance des mots ambigus
calcul de similarité sémantique
Évaluation WSD/sélection lexicale

Annexe B

Les algorithmes présentés dans cet Annexe font partie d'un ensemble de programmes et de scripts développés en Perl (version 5.8.8) sous Linux Ubuntu 6.10 (Edgy Eft). L'ensemble des scripts de transformation des données (corpus) et des programmes comporte entre 12.000 et 15.000 lignes de code incluant les commentaires. L'ensemble est disponible sur simple demande auprès de l'auteur.

Algorithmes présents dans le Chapitre 6

Typologie des éléments utilisés par les algorithmes présentés

L'implémentation est structurée de manière à avoir un accès rapide aux propriétés des données manipulées, via plusieurs tables associatives, dont les clés sont soit des mots ambigus, soit des clusters, soit des EQVs individuels, soit des paires d'EQVs. Nous décrivons, ci-après, les types des données utilisées ainsi que leurs propriétés.

- **mot_ambigu(M)** : mot : *chaîne*
- **trait** : mot : *chaîne*
- **poids** : *flottant*
- **seuil** : *flottant*
- **EQV** : mot : *chaîne*
  -> liste_de_traits_LS (Langue Source) et leurs poids (liste d'éléments pondérés)
  -> liste_de_traits_LC (Langue Cible) et leurs poids (liste d'éléments pondérés)
- **paire_EQVs** : EQV1 : EQV
  EQV2 : EQV
  -> traits_assimilateurs (EQV1, EQV2) :

- traits_assimil_LS ← liste_de_traits_LS_EQV1 ∩ liste_de_traits_LS_EQV2

- traits_assimil_LC ← liste_de_traits_LC_EQV1 ∩ liste_de_traits_LC_EQV2

- **table_sim** : tableau à deux dimensions dont les indices sont des mots ; la valeur de chaque cellule est calculée par la mesure de Jaccard pondérée (cf. §2.3.2, chapitre 6) (table de similarité)

- **cluster** : liste d'EQVs

  -> traits (LS) qui caractérisent chaque cluster

Algorithme Principal de Clustering du Programme SEMCLU

Dans cette section, nous décrivons en détail le fonctionnement de l'algorithme de clustering. Nous reproduisons la description donnée dans le paragraphe 2.5 du chapitre 6 afin de faciliter la compréhension du pseudocode, et nous fournissons davantage de détails concernant l'implémentation.

***Objectifs* :** Renvoyer une liste contenant une association entre chaque **mot ambigu** analysé et la **liste des clusters d'EQVs** générés par SEMCLU pour ce mot.

***Etat* :** Une **table de similarité** a été créée pour chaque mot ambigu (cf. §2.3, chapitre 6), contenant les scores de similarité entre les paires de ses EQVs (par rapport aux contextes de la LS).

Description du fonctionnement de l'algorithme présenté en paragraphe 2.5 du chapitre 6 :

1. L'algorithme **prend en entrée** :
   - la **liste des EQVs** d'un mot ambigu, trouvés dans le lexique bilingue
   - la **table de similarité** du mot (dans la LS)
   - le **seuil**, qui correspond à la moyenne des scores attribués aux paires des EQVs du mot, sans considération des paires dont le score est égal à 0.
2. La liste des EQVs, la table de similarité et le seuil constituent l'entrée de la fonction ***création_clusters_initiaux***, qui génère les '**bons'** et les '**mauvais clusters'** à deux éléments. Cette fonction renvoie également la partie de la table de similarité du mot M qui contient les paires ayant un score > seuil (moyenne).
3. La liste de 'bons clusters' issue de *création_clusters_initiaux* contient des doublons (étant donné que la relation entre chaque paire d'EQVs est

trouvée deux fois dans la table, par ex. table[EQV1][EQV2] et table [EQV2][EQV1]). Les doublons sont éliminés afin de ne retenir chaque cluster qu'une seule fois.

4. Les mots inclus dans de 'bons clusters' sont mis dans une liste plate, sans doublons ('liste_EQVs_clustérisés').
5. Chaque **bon cluster** ('Cluster_C') constitue l'entrée de la fonction ***enrichissement_cluster***, qui entreprend de l'enrichir par d'autres éléments. Pour ce faire, elle se sert de la liste de tous les mots contenus dans de bons clusters ('listeEQVs_clustérisés'). La fonction *enrichissement_cluster* renvoie chaque cluster éventuellement enrichi par des EQVs ayant une relation significative avec ceux déjà contenus dans le cluster d'entrée. A noter que, étant donné que le cluster d'entrée contient uniquement 2 EQVs, chaque EQV candidat pour être inclus dans le cluster est comparé **uniquement** avec ces deux EQVs. Dans le cas donc où d'autres EQVs ont été inclus, par la fonction *enrichissement _cluster*, dans le cluster avant lui, il se peut que ces EQVs n'aient pas de relation significative avec lui (l'existence d'une telle relation n'est pas vérifiée). Le cluster généré par la fonction *enrichissement_cluster* est donc considéré comme un **cluster temporaire**. Autre élément important : l'enrichissement des 'bons clusters' peut résulter à leur fusion (cf. §2.5.3, chapitre 6) ou à leur chevauchement (cf. §2.5.4, chapitre 6).
6. Chaque cluster temporaire issu de la fonction *enrichissement_cluster* constitue l'entrée de la fonction ***purge_cluster***. Cette fonction entreprend, quant à elle, de supprimer du cluster temporaire les éléments n'entretenant pas de relations avec **TOUS** les autres éléments du cluster, sachant que tous les éléments d'un cluster final doivent être liés de manière significative entre eux. Lorsque cette condition est satisfaite et qu'il ne reste plus d'éléments dans le cluster qui ne sont pas liés avec tous les autres, *purge* s'arrête.

7. La liste des clusters enrichis et nettoyés peut contenir des doublons ; ces doublons sont éliminés pour ne retenir chaque cluster qu'une seule fois dans la liste.
8. La prochaine étape consiste à créer une liste des éléments des 'mauvais clusters' ('listeM_Cm') qui n'ont pas été inclus dans AUCUN des clusters enrichis et nettoyés. La liste plate des éléments inclus dans les clusters en question coïncide avec la 'liste_EQVs_clustérisés' (qui contient l'ensemble des éléments des 'bons clusters' initialement créés). Pour trouver les éléments de la 'listeM_Cm' qui n'ont pas été inclus dans aucun des clusters précédemment générés, nous calculons la **complémentaire** de 'liste_EQVs_clustérisés' dans 'listeM_Cm' (notée 'C_{listeM_Cm}liste_EQVs_clustérisés') ; la complémentaire correspond ici à l'ensemble formé des éléments de 'listeM_Cm' qui **ne sont pas** dans 'liste_EQVs_clustérisés'. Autrement dit, il s'agit de l'ensemble des éléments de 'listeM_Cm' qui lui sont uniques, c'est-à-dire qui n'appartiennent pas à son intersection avec la 'liste_EQVs_clustérisés' (si les deux ensembles ont des éléments en commun).
9. Chaque EQV de la liste générée pendant l'étape précédente est, ensuite, inclus dans un cluster ne contenant QUE cet EQV ; les clusters qui apparaissent plus d'une fois dans cette liste sont éliminés.
10. La liste '**clusters_finaux'** du mot ambigu (M) contient, d'une part, les clusters enrichis et nettoyés (clusters_enrichis_nettoyés) et, d'autre part, les clusters générés pendant l'étape 9, dont chacun ne contient qu'un seul EQV.

Outre la liste des clusters de chaque mot ambigu, à la sortie du programme SEMCLU nous avons accès à d'autres éléments (i.e. les traits de la LS caractérisant les clusters) qui seront exploités lors des processus de désambiguïsation et de sélection lexicale.

Description de l'algorithme en pseudocode :

Paramètres :

listeEQVs_M	// liste des EQVs d'un mot_ambigu M
table_sim	// table de similarité des EQVs du mot_ambigu M dans la LS
seuil	// seuil à partir duquel on considère que deux EQVs sont liés par une relation significative (= la moyenne des scores de la table de similarité, sans considération des paires avec score égal à 0)

Variables :

liste_C_Bons	// liste de clusters
liste_C_Mauvais	// liste de clusters
liste_EQVs_clustérisés	// liste d'EQVs
cluster_C	// cluster
cluster_Cm	// cluster
temp_cluster	// cluster
EQVs_non_clustérisés	// liste d'EQVs
listeM_Cm	// liste d'EQVs
liste_clusters_un_élément	// liste de clusters
clusters_enrichis_nettoyés	// liste de clusters
liste_clusters_finaux	// liste de clusters
table_sim_sup_moyenne	// table de similarité

Retourne *:*

liste_cluster_finaux	// liste de clusters du mot M

(Note : le nombre précédé d'un commentaire renvoie à l'explication donnée page précédente)

Etant donné un mot ambigu M

DEBUT

// Initialisation
liste_C_Bons ← Ø
liste_C_Mauvais ← Ø
table_sim_sup_moyenne ← Ø
listeEQVs_clustérisés ← Ø
liste_clusters_finaux ← Ø

{liste_C_Bons, liste_C_Mauvais, table_sim_sup_moyenne} ← *création_clusters_initiaux*(listeEQVs_M, table_sim, seuil) // 2

liste_C_Bons ← *renvoie_clusters_sans_doublons*(liste_C_Bons) // 3

listeEQVs_clustérisés ← *renvoie_EQVs_clustérisés_sans_doublons*(liste_C_Bons) // 4

clusters_enrichis_nettoyés ← Ø

POUR chaque cluster_C de liste_C_Bons **FAIRE**
    temp_cluster ← *enrichissement_cluster*(cluster_C, liste_EQVs_clustérisés) // 5
    temp_cluster ← *purge_cluster*(cluster_C) // 6
    clusters_enrichis_nettoyés ← clusters_enrichis_nettoyés U {temp_cluster}
FINPOUR

clusters_enrichis_nettoyés ← *renvoie_liste_sans_doublons(*clusters_ enrichis_nettoyés) // 7

EQVs_non_clustérisés ← Ø
POUR Chaque cluster_Cm de liste_C_Mauvais **FAIRE**
    listeM_Cm ← *renvoie_liste_EQVs*(cluster_Cm)
    EQVs_non_clustérisés ← EQVs_non_clustérisés U C_{listeM_Cm}liste_EQVs_clustérisés // 8
FINPOUR

liste_clusters_un_élément ← *crée_clusters_sans_doublons*(EQVs_non_clustérisés) // 9

liste_clusters_finaux ← liste_clusters_un_élément U clusters_enrichis_nettoyés // 10

RETOURNE liste_clusters_finaux
FIN

Fonctions Présentes dans l'Algorithme de Clustering du Programme SEMCLU

Liste des fonctions :

- *création_clusters_initiaux*
- *enrichissement_cluster*
- *purge_cluster*

création_clusters_initiaux

Objectifs : A partir de la liste des équivalents (EQVs) et de la table de similarité d'un mot M, générer les listes de 'bons' et de 'mauvais' clusters. Les 'bons' contiennent les paires d'EQVs ayant un score supérieur à la moyenne, tandis que les 'mauvais' contiennent les paires ayant un score inférieur à la moyenne. Les clusters formés par cette fonction peuvent se chevaucher ; chevauchement qui est décrit par l'intersection non vide des clusters. La fonction retourne également la partie de la table de similarité contenant les paires d'EQVs ayant un score supérieur à la moyenne. La sortie de cette fonction servira, par la suite, à la création des clusters finaux (par la fonction *enrichissement_cluster*).

Paramètres :

listeEQVs_M // liste des EQVs d'un mot M
table_sim // table de similarité des EQVs du mot_ambigu M dans la LS
seuil // seuil à partir duquel on considère que deux EQVs sont liés par une relation significative

Variables :

liste_C_Bons // liste de clusters
liste_C_Mauvais // liste de clusters
listeEQVs_M // liste d'EQVs
EQV1, EQV2 // EQVs
seuil, score // flottants

Retourne :

{liste_C_Bons, liste_C_Mauvais, table_sim_sup_moyenne} avec
liste_C_Bons // liste clusters à 2 EQVs retenant des relations significatives
liste_C_Mauvais // liste clusters à 2 EQVs ne retenant pas de relation significative
table_sim_sup_moyenne // table contenant les paires d'EQVs ayant une relation significative et leur score

Note : Pour faciliter la lecture, la table de similarité est directement indexée par des mots au lieu d'indices. Dans l'implémenation il s'agit d'une table associative à deux clefs.

```
DEBUT
    // Initialisation
    liste_C_Mauvais← Ø
    liste_C_Bons← Ø
    table_sim_sup_moyenne← Ø

    POUR chaque paire<EQV1, EQV2> de listeEQVs_M FAIRE
        score = table_sim[EQV1]-[EQV2]
        SI score > seuil ALORS
            liste_C_Bons ← liste_C_Bons U {EQV1, EQV2}
            table_sim_sup_moyenne ← table_sim_sup_moyenne  U {EQV1-EQV2, score}
        SINON
            liste_C_Mauvais ← liste_C_Mauvais U {EQV1, EQV2}
        FINSI
    FINPOUR
    RETOURNE (liste_C_Bons, liste_C_Mauvais, table_sim_sup_moyenne)
FIN
```

enrichissement_cluster

Objectifs **:** Parcourir les clusters initiaux ('liste_C_bons') et ajouter d'autres EQVs à ces clusters, si possible. Les clusters formés par cette fonction peuvent se chevaucher, comme c'était également le cas pour les 'bons clusters' ('C_Bons', issus de la fonction *creation_clusters_initiaux*). Le chevauchement des clusters est décrit, ici aussi, par leur intersection non vide.
On définit la fonction *grep* comme renvoyant les associations entre un EQV et un élément d'un cluster à partir de la liste des associations de l'EQV, incluses dans la table de similarité.

Etat **:** Chaque cluster traité vient d'être créé par la fonction *création_clusters_initiaux* et, par conséquent, il ne contient que 2 éléments.

Paramètres **:**

CLU	// cluster issu de C_Bons
liste_ EQVs_clustérisés	// liste des EQVs clustérisés par la fonction *creation_clusters_initiaux*, issue de C_Bons, sans doublons
table_sim_sup_moyenne	// table de similarité issue de *création_clusters_initiaux*

Variables :

CLU	// liste d'EQVs
EQV, EQV1, EQV2	// EQVs
liste_assoc_EQV	// liste de paires d'EQVs
liste_assoc_EQV1	// liste de paires d'EQVs
liste_assoc_EQV2	// liste de paires d'EQVs

Retourne :

CLU	// cluster éventuellement enrichi

Etant donné un cluster CLU

DEBUT

{EQV1, EQV2} ← *renvoie_éléments*(CLU) // on récupère le cluster et on met ses éléments dans une liste

POUR chaque EQV **de** EQVs_clustérisés **FAIRE**
liste_assoc_EQV ← Ø
SI EQV **NON** ∈ CLU **FAIRE**
liste_assoc_EQV ← renvoie_assoc (EQV, *table_similarité_sup_moyenne*)
liste_assoc_EQV1 ← grep(EQV1, liste_assoc_EQV)
liste_assoc_EQV2 ← grep(EQV2, liste_assoc_EQV)
SI (liste_assoc_EQV1 non vide) **ET** (liste_assoc_EQV2 non vide) **FAIRE**
CLU ← CLU U {EQV}
FINSI
FINSI
FINPOUR
RETOURNE CLU
FIN

purge_cluster

Objectifs : Retirer d'un cluster les éléments qui n'ont pas de relations significatives (i.e. score de similarité supérieur à la moyenne) avec TOUS les autres éléments du cluster.

Etat : Un cluster vient d'être traité par la fonction *enrichissement_cluster* et contient éventuellement des EQVs n'ayant pas de relations avec tous les autres EQVs inclus dans le cluster.

Paramètres :

CLU	// cluster issu de *enrichissement_cluster*
table_sim_sup_moyenne	// table issue de *création_clusters_initiaux*

Variables :

liste_éléments_cluster	// liste d'EQVs
liste_relations_significatives_EQV	// liste de paires d'EQVs
CLU	// liste d'EQVs
EQV	// EQV

Retourne :

CLU	// cluster contenant des éléments étant tous liés entre eux (dont les éléments non liés à tous les autres ont été éliminés)

Etant donné un cluster CLU

DEBUT
liste_éléments_cluster ← *renvoie_liste_éléments*(CLU)
POUR chaque EQV_x de liste_éléments_cluster **FAIRE**
liste_relations_significatives_EQV_x ← *renvoie_relations (*EQV_x, table_sim_sup_moyenne)
POUR chaque EQV_y **de** liste_éléments_cluster
SI $EQV_x \neq EQV_y$ **FAIRE**
SI {EQV_x, EQV_y} **NON** ∈ liste_relations_significatives_EQV_x
CLU ← CLU – {EQV_y}
FINSI
FINSI
FINPOUR
FINPOUR
RETOURNE CLU
FIN

Annexe C

Algorithmes présents dans le Chapitre 5

Typologie des données utilisées par les algorithmes :

SEUIL_fréquence : *flottant*

SEUIL_cooccurrence : *flottant*

SEUIL_nbre_voisins_directs : *entier*

mot : mot : *chaîne*

poids : *flottant*

N (nœud) : mot : *chaîne*

A (arête) : <nœud$_x$, nœud$_y$, poids>

G : ensemble de nœuds et d'arêtes

graphe_EQV : graphe de type G

liste_mot_graphes : liste de couples <mot, G>

liste_fréq : liste de mots et leur fréquence (<mot, fréq>)

nbre_voisins_directs : entier

liste_voisins_directs : liste de mots

C : matrice de cooccurrence : table à deux dimensions dont les indices sont des mots ; la valeur de chaque cellule correspond à la fréquence de cooccurrence de deux mots

liste_assoc_EQV : liste de couples < nœud$_x$, nœud$_y$> (i.e. arêtes sans les poids)

liste_assoc_G : liste de couples < nœud$_x$, nœud$_y$> (i.e. arêtes sans les poids)

liste_assoc_pondérées : liste d'arêtes A d'un graphe G

assoc_communes : liste d'arêtes A

Liste des algorithmes :

- Algorithme de construction des graphes de cooccurrence correspondants à un mot ambigu

- Algorithme de calcul de recouvrement entre le graphe d'un EQV et les graphes d'un mot ambigu

Algorithme de construction des graphes de cooccurrence correspondants à un mot ambigu

Description de l'algorithme présenté dans le paragraphe 1.2.3. du chapitre 5 :

- L'algorithme sélectionne le mot le plus fréquent (mot_{Fmax}) de la liste de fréquence construite pour un mot ambigu (M).
- Si la fréquence de mot_{Fmax} est supérieure à un seuil ('SEUIL_fréquence'), il construit une matrice de cooccurrence (C_{fmax}) pour ce mot.
- Si mot_{Fmax} a une fréquence de cooccurrence supérieure à un seuil ('SEUIL_cooccurrence') avec un nombre suffisant de mots (voisins directs) (nombre supérieur au 'SEUIL_nombre_voisins_directs'), un graphe ('G') est construit pour mot_{Fmax}. Les nœuds du graphe sont les mots ayant une fréquence de cooccurrence significative avec mot_{Fmax} (cf. §1.2.3, chapitre 5). L'arête liant deux nœuds est pondérée par la formule 1.2.3.1 (chapitre 5) (fonction *calcule_poids*).
- Une fois le graphe de mot_{Fmax} construit, mot_{Fmax} et ses voisins dans le graphe sont éliminés de la liste de fréquence de M.
- Le mot le plus fréquent de la liste de fréquence résultante est sélectionné et la procédure continue de la même manière. Ce processus est réitéré tant que la fréquence du mot le plus fréquent est supérieure au 'SEUIL_fréquence'.

Description de l'algorithme en pseudocode:

Paramètres :

SEUIL_fréquence	// flottant
SEUIL_cooccurrence	// flottant
SEUIL_nbre_voisins_directs	// entier

Variables :

liste_mot_graphes	// liste de couples <mot, G>
liste_fréq	// liste de mots et leur fréquence
nbre_voisins_directs	// entier
liste_voisins_directs	// liste de mots
mot	// mot
C	// matrice de cooccurrence
G	// ensemble de nœuds et d'arêtes
liste_N	// liste de nœuds
liste_A	// liste d'arêtes
poids	// flottant

Retourne :

liste_mot_graphes	// liste de couples <mot, G>

Etant donné un mot ambigu M

DEBUT

// Initialisation
liste_mot_graphes ← Ø
liste_fréq = *renvoie_liste_frequence_mot_ambigu*(M)

TANT QUE *freq*(*renvoie_couple_mot_fréquence_max*(liste_fréq)) > SEUIL_fréquence **FAIRE**
mot$_{Fmax}$ ← *mot*(*renvoie_couple_mot_fréquence_max*(liste_fréq))
C$_{fmax}$ = *construire_matriceCooccurrence*(mot$_{Fmax}$)
nbre_voisins_directs ← 0
liste_voisins_directs ← Ø
POUR chaque mot$_j$ de C$_{fmax}$ [mot$_{Fmax}$] **FAIRE**
SI C$_{fmax}$ [mot$_{Fmax}$][mot$_j$] > SEUIL_cooccurrence **ALORS**
nbre_voisins_directs++
liste_voisins_directs ← liste_voisins_directs U {mot$_j$}
FINSI
FINPOUR
SI nbre_voisins_directs > SEUIL_nbre_voisins_directs **ALORS**
G$_{fmax}$ ← Ø
liste_N$_{fmax}$ ← Ø
liste_A$_{fmax}$ ← Ø
POUR chaque mot$_i$ et mot$_j$ **de** C$_{fmax}$ **FAIRE**
SI C$_{fmax}$ [Mot$_i$][Mot$_j$] > SEUIL_cooccurrence **ALORS**
poids ← 0
liste_N$_{fmax}$ ← liste_N$_{fmax}$ U {Mot$_i$}

liste_N_{fmax} ← liste_N_{fmax} U {Mot_j}
poids(Mot_i, Mot_j) = *calcule_poids*(Mot_i, Mot_j)
liste_A_{fmax} ← liste_A_{fmax} U {Mot_i, Mot_j, poids(Mot_i, Mot_j)}
G_{fmax} ← G_{fmax} U {liste_N_{fmax}, liste_A_{fmax}}

FINSI

FINPOUR

FINSI

liste_mot_graphes ← liste_mot_graphes U {mot_{Fmax}, G_{fmax}}
liste_fréq ← liste_fréq – mot_{fmax}
liste_fréq ← liste_fréq – liste_voisins_directs

FINTQ

RETOURNE liste_mot_graphes

FIN

Algorithme de calcul de recouvrement
entre le graphe d'un EQV et les graphes d'un mot ambigu

Description du fonctionnement de l'algorithme :

- L'algorithme prend en entrée le graphe d'un EQV (graphe_EQV).
- Il renvoie la liste des associations présentes dans 'graphe_EQV' avec et sans les poids qui leur sont attribués.
- Pour chaque graphe source 'G', il récupère la liste des associations de 'G'.
- Il trouve l'intersection de la liste d'associations de 'G' et de la liste d'associations de 'graphe_EQV'.
- Si l'intersection n'est pas vide, il récupère le poids que chaque association de l'intersection a dans le 'graphe_EQV'.
- L'algorithme retourne l'intersection des associations des graphes 'graphe_EQV' et 'G' contenant leur poids.

On définit la fonction *renvoie_assoc* comme renvoyant les associations entre deux nœuds d'un graphe et la fonction *renvoie_assoc_avec_poids* comme renvoyant les associations et les poids qui leur sont associés (i.e. le poids des arêtes liant les nœuds dans le graphe). La fonction *renvoie_poids* renvoie, quand à elle, le poids d'une association.

Description en pseudocode de l'algorithme présenté en paragraphe 2.4.1. du chapitre 5.

Paramètres :

graphe_EQV	// ensemble de nœuds et d'arêtes

Variables :

association	// couple < $nœud_x$, $nœud_y$>
liste_assoc_G	// liste de couples < $nœud_x$, $nœud_y$>
liste_assoc_EQV	// liste de couples < $nœud_x$, $nœud_y$>
liste_intersection	// liste de couples < $nœud_x$, $nœud_y$>
liste_intersection_pondérée	// liste d'arêtes
liste_assoc_EQV_pondérées	// liste d'arêtes
G	// ensemble de nœuds et d'arêtes
poids	// flottant

Retourne :

assoc_communes <graphe_EQV, G>	// liste d'associations communes entre les deux graphes

```
assoc_communes ← Ø
liste_assoc_EQV ← Ø
liste_assoc_EQV_pondérées ← Ø

Etant donné un graphe_EQV

DEBUT
    //Initialisation
    liste_assoc_EQV ← renvoie_assoc(graphe_EQV)
    liste_assoc_EQV_pondérées ← renvoie_assoc_avec_poids(graphe_EQV)
    liste_intersection ← Ø
    liste_intersection_pondérée ← Ø

    POUR chaque graphe source G
        liste_assoc_G ← renvoie_assoc(G)
        liste_intersection ← liste_assoc_G ∩ liste_assoc_EQV
        SI liste_intersection non vide
            POUR chaque association de liste_intersection
                poids ← renvoie_poids(association, liste_assoc_EQV_pondérées)
                liste_intersection_pondérée ← liste_intersection_pondérée U <association, poids>
            FINPOUR
            assoc_communes <graphe_EQV, G> ← liste_intersection_pondérée
        FINSI
    FINPOUR

    RETOURNE assoc_communes <graphe_EQV, G>
FIN
```

Annexe D

D.1 Entrées retenues du lexique bilingue anglais-grec automatiquement généré

1. **accommodation** : φιλοξενία στέγαση κατάλυμα
2. **achievement** : επίτευξη επίδοση υλοποίηση επίτευγμα
3. **adequacy** : ορθότητα επάρκεια καταλληλότητα
4. **administration** : διαχείριση χορήγηση διοίκηση
5. **admission** : είσοδος εισδοχή εγγραφή εισαγωγή
6. **adoption** : έκδοση υιοθέτηση έγκριση θέσπιση λήψη
7. **alcohol** : αλκοόλη οινόπνευμα αλκοόλ
8. **amendment** : τροποποίηση αναθεώρηση τροπολογία
9. **area** : τομεύς έκταση πεδίο χώρος περιοχή
10. **argument** : συλλογισμός ισχυρισμός επιχειρηματολογία επιχείρημα
11. **arrangement** : ρύθμιση λεπτομέρεια διακανονισμός διευθέτηση μηχανισμός καθεστώς
12. **assembly** : συνέρχεσθαι συνέλευση συνεταιρίζεσθαι
13. **assistance** : βοήθεια ενίσχυση συνδρομή αρωγή
14. **association** : σύλλογος συμμετοχή σύνδεση οργάνωση συνεταιρίζεσθαι ένωση
15. **attachment** : προσάρτημα προσήλωση κατάσχεση
16. **attitude** : συμπεριφορά αντίληψη στάση
17. **awareness** : ευαισθητοποίηση συνειδητοποίηση συνείδηση
18. **beach** : ακτή παραλία αμμουδιά
19. **belief** : πεποίθηση πίστη δοξασία
20. **bill** : τιμολόγιο νομοσχέδιο λογαριασμός
21. **car** : όχημα αυτοκίνητο αμάξι
22. **care** : φροντίδα περίθαλψη μέριμνα
23. **change** : αλλαγή μεταβολή τροποποίηση
24. **channel** : οδός κανάλι δίαυλος
25. **charge** : τέλος επιβάρυνση φόρος τιμολόγηση
26. **church** : εκκλησία ναός εκκλησιά
27. **circumstance** : συνθήκη περίσταση περίπτωση
28. **clothing** : ένδυση ιματισμός ρούχο
29. **coating** : επίστρωση επίχρισμα επίχριση
30. **communication** : κοινοποίηση επικοινωνία ανακοίνωση
31. **complaint** : καταγγελία αναφορά ένσταση
32. **completion** : συμπλήρωση ολοκλήρωση περάτωση

33. **compliance** : τήρηση σεβασμός συμμόρφωση
34. **conclusion** : πόρισμα σύναψη συμπέρασμα
35. **conference** : συνδιάσκεψη συνέδριο διάσκεψη
36. **consent** : συμφωνία συγκατάθεση έγκριση συναίνεση
37. **consultation** : διάλογος διαβούλευση συνεννόηση
38. **contribution** : συμβολή συνεισφορά εισήγηση συμμετοχή εισφορά
39. **controversy** : διαμάχη αντιδικία αντιπαράθεση
40. **defect** : ελάττωμα σφάλμα ανωμαλία
41. **definition** : καθορισμός ορισμός προσδιορισμός
42. **department** : υπουργείο τμήμα υπηρεσία εξουσία νομός
43. **designation** : χαρακτηρισμός διορισμός ονομασία ορισμός
44. **detection** : ανίχνευση διάγνωση εξακρίβωση
45. **determination** : διαπίστωση βούληση αποφασιστικότητα καθορισμός προσδιορισμός
46. **difference** : διαφορά απόκλιση διαφορετικότητα
47. **disease** : νόσημα νόσος ασθένεια πάθηση
48. **distribution** : κατανομή διανομή διάδοση
49. **disturbance** : παρακώλυση διατάραξη διαταραχή
50. **diversity** : ποικιλία πολυμορφία ποικιλότητα ποικιλομορφία
51. **division** : περιφέρεια κατανομή διαχωρισμός διαίρεση
52. **duty** : υποχρέωση δασμός φόρος καθήκον
53. **education** : κατάρτιση παιδεία εκπαίδευση αγωγή
54. **efficiency** : απόδοση αποδοτικότητα αποτελεσματικότητα
55. **elimination** : εξάλειψη αποβολή κατάργηση
56. **equivalence** : αντιστοιχία ισοτιμία ισοδυναμία
57. **error** : ελάττωμα λάθος σφάλμα πλάνη
58. **establishment** : ίδρυση δημιουργία εγκατάσταση καθιέρωση κέντρο θέσπιση ίδρυμα
59. **event** : εκδήλωση περιστατικό συμβάν γεγονός περίπτωση
60. **facility** : υποδομή εγκατάσταση διευκόλυνση εξοπλισμός
61. **fax** : τηλεομοιοτυπία τέλεφαξ φαξ
62. **feast** : γιορτή εκδήλωση πανηγύρι
63. **film** : φιλμ ταινία επίστρωση μεμβράνη
64. **finding** : πόρισμα διαπίστωση εκτίμηση εύρημα
65. **foot** : πόδας πρόποδας πόδι
66. **formulation** : χάραξη διατύπωση διαμόρφωση σύνθεση
67. **function** : αποστολή καθήκον λειτουργία
68. **guardian** : κηδεμών θεματοφύλακας επίτροπος κηδεμόνας
69. **guidance** : συμβουλή προσανατολισμός καθοδήγηση
70. **identification** : ταυτότητα εντοπισμός προσδιορισμός αναγνώριση ταυτοποίηση
71. **implementation** : εκτέλεση εφαρμογή υλοποίηση
72. **implication** : συνέπεια επιπλοκή επίπτωση
73. **information** : πληροφόρηση στοιχείο πληροφορία ενημέρωση
74. **injury** : τραύμα τραυματισμός βλάβη

75. **institution** : όργανο φορέας οργανισμός θεσμός ίδρυμα
76. **integration** : ολοκλήρωση ένταξη ενσωμάτωση ενοποίηση
77. **interest** : ενδιαφέρον τόκος συμφέρον ωφέλεια
78. **introduction** : καθιέρωση εισαγωγή κίνηση θέσπιση
79. **inventory** : απογραφή απολογισμός κατάλογος ευρετήριο
80. **issue** : έκδοση πρόβλημα θέμα ζήτημα
81. **label** : σήμα επισήμανση ετικέτα χαρτόνι σήμανση
82. **line** : άξονας κατεύθυνση γραμμή
83. **maintenance** : συντήρηση διατροφή διατήρηση
84. **management** : διαχείριση διεύθυνση διοίκηση
85. **manager** : διευθυντής στέλεχος διαχειριστής
86. **manufacture** : κατασκευή παραγωγή παρασκευή
87. **manufacturing** : κατασκευή παραγωγή παρασκευή
88. **meal** : γεύμα εστίαση τροφή
89. **model** : μοντέλο πρότυπο υπόδειγμα
90. **momentum** : ώθηση ταχύτητα ορμή
91. **motivation** : παρακίνηση ενθάρρυνση κίνητρο δραστηριοποίηση
92. **movement** : μετακίνηση διακίνηση κίνηση κυκλοφορία
93. **multitude** : πλήθος πολλαπλότητα πληθώρα
94. **name** : ονομασία επωνυμία όνομα
95. **nationality** : ιθαγένεια υπηκοότητα εθνικότητα
96. **opening** : διάνοιξη απελευθέρωση άνοιγμα
97. **organisation** : οργανισμός διοργάνωση οργάνωση
98. **outcome** : πόρισμα έκβαση αποτέλεσμα
99. **overview** : εικών ανασκόπηση επισκόπηση
100. **paper** : βίβλος έγγραφο βιβλίο
101. **passage** : χωρίο πέρασμα πάροδος
102. **penetration** : εισχώρηση διάτρηση διείσδυση
103. **period** : προθεσμία περίοδος διάστημα διάρκεια
104. **player** : φορέας παράγοντας πρωταγωνιστής συντελεστής
105. **power** : ισχύς δύναμη εξουσία αρμοδιότητα
106. **preparation** : εκπόνηση προετοιμασία παρασκεύασμα σκεύασμα παρασκευή
107. **preservation** : διάσωση διατήρηση διαφύλαξη
108. **promotion** : προβολή προαγωγή προώθηση
109. **prosperity** : ακμή ευμάρεια ευημερία
110. **question** : ζήτημα ερώτημα πρόβλημα θέμα ερώτηση
111. **range** : εύρος ποικιλία φάσμα σειρά
112. **record** : αρχείο φάκελος μητρώο καταγραφή
113. **reference** : μνεία παραπομπή αναφορά
114. **region** : περιφέρεια περιοχή διαμέρισμα
115. **registration** : καταχώρηση καταχώριση καταγραφή εγγραφή
116. **release** : απελευθέρωση αποδέσμευση ελευθέρωση
117. **relevance** : σημασία συνάφεια ενδιαφέρον σπουδαιότητα καταλληλότητα

118. **removal** : εξάλειψη αφαίρεση απομάκρυνση
119. **renovation** : ανακαίνιση αναπαλαίωση ανανέωση
120. **reorganisation** : αναμόρφωση αναδιαμόρφωση αναδιοργάνωση
121. **replacement** : ανανέωση αναπλήρωση αντικατάσταση
122. **response** : αντίδραση απάντηση ανταπόκριση
123. **responsibility** : ευθύνη μέριμνα αρμοδιότητα
124. **restoration** : αποκατάσταση ανάκτηση αναζωογόνηση συντήρηση αναστήλωση
125. **rock** : βράχος πέτρωμα πέτρα
126. **route** : οδός διαδρομή δρομολόγιο δρόμος
127. **scope** : έκταση εμβέλεια εύρος πεδίο
128. **sector** : τομεάς, κλάδος, τομεύς
129. **section** : τμήμα ενότητα κεφάλαιο μέρος
130. **session** : συνεδρίαση συνεδρία σύνοδος
131. **settlement** : συμβιβασμός οικισμός διευθέτηση διακανονισμός
132. **shortcoming** : κενό έλλειψη ανεπάρκεια
133. **specialisation** : εξειδίκευση ειδικότητα ειδίκευση
134. **statement** : δήλωση δικόγραφο υπόμνημα
135. **street** : δρομάκι δρόμος σοκάκι
136. **structure** : κατασκευή διάρθρωση δομή
137. **student** : σπουδαστής φοιτητής μαθητής
138. **suggestion** : εισήγηση πρόταση υπόδειξη
139. **supplement** : συμπλήρωση συμπλήρωμα προσθήκη
140. **surveillance** : επιτήρηση παρακολούθηση εποπτεία επίβλεψη
141. **survey** : δημοσκόπηση έρευνα επισκόπηση
142. **time** : προθεσμία διάστημα φορά χρόνος στιγμή
143. **trade** : επάγγελμα συναλλαγή εμπόριο
144. **training** : επιμόρφωση κατάρτιση εκπαίδευση
145. **treatment** : επεξεργασία αντιμετώπιση θεραπευτική θεραπεία αγωγή μεταχείριση
146. **uncertainty** : ανασφάλεια αβεβαιότητα ασάφεια αμφιβολία
147. **variation** : μεταβολή τροποποίηση διακύμανση
148. **waste** : απόρριμμα λύμα απόβλητο
149. **willingness** : βούληση επιθυμία προθυμία
150. **withdrawal** : ανάκληση αποχώρηση υπαναχώρηση αφαίρεση απόσυρση

D.2 Taille de l'échantillon de test des mots du lexique bilingue anglais-grec automatiquement généré (donnée en nombre d'instances de test)

department : 298
υπουργείο 13
υπηρεσία 211
τμήμα 65
νομός 9

clothing : 105
ένδυση 98
ρούχο 7

complaint : 350
αναφορά 16
καταγγελία 330
ένσταση 4

meal : 20
γεύμα 14
τροφή 6

elimination : 160
κατάργηση 84
εξάλειψη 76

passage : 17
πέρασμα 10
πάροδος 7

opening : 245
άνοιγμα 240
διάνοιξη 2
απελευθέρωση 3

preparation : 890
παρασκεύασμα 66
εκπόνηση 32
σκεύασμα 14
παρασκευή 10
προετοιμασία 768

conclusion : 2357
σύναψη 167
πόρισμα 147
συμπέρασμα 2043

issue : 11078
πρόβλημα 948
έκδοση 38
θέμα 6260
ζήτημα 3832

implication : 571
συνέπεια 221
επιπλοκή 8
επίπτωση 342

achievement : 368
επίτευξη 99
επίδοση 24
υλοποίηση 48
επίτευγμα 197

statement : 2276
δήλωση 2276

implementation : 3153
εκτέλεση 458
εφαρμογή 2057
υλοποίηση 638

foot : 54
πόδι 54

power : 3768
ισχύς 237
εξουσία 1793
δύναμη 740
αρμοδιότητα 998

reorganisation : 45
αναμόρφωση 1
αναδιοργάνωση 44

withdrawal : 245
αποχώρηση 57
ανάκληση 10
απόσυρση 148
υπαναχώρηση 12
αφαίρεση 18

compliance : 460
τήρηση 223
σεβασμός 59
συμμόρφωση 178

period : 2739
διάρκεια 137
προθεσμία 217
περίοδος 2039
διάστημα 346

supplement : 87
συμπλήρωμα 77
προσθήκη 3
συμπλήρωση 7

renovation : 11
ανακαίνιση 6
ανανέωση 5

name : 1090
ονομασία 125
επωνυμία 2
όνομα 963

interest : 6232
συμφέρον 4581
ωφέλεια 107
ενδιαφέρον 1377
τόκος 167

distribution : 727
διανομή 341
διάδοση 33
κατανομή 353

alcohol : 303
οινόπνευμα 145
αλκοόλη 9
αλκοόλ 149

introduction : 1138
κίνηση 2
εισαγωγή 629
καθιέρωση 349
θέσπιση 158

consultation : 1242
διάλογος 14
διαβούλευση 1191
συνεννόηση 37

multitude : 19
πληθώρα 11
πλήθος 6
πολλαπλότητα 2

feast : 1
γιορτή 1

church : 124
εκκλησία 123
ναός 1
εκκλησιά 0

time : 5875
φορά 2704
χρόνος 29
προθεσμία 338
στιγμή 2208
διάστημα 596

penetration : 24
διείσδυση 24

area : 9099
τομεύς 5429
έκταση 168
πεδίο 304
χώρος 23
περιοχή 3175

argument : 1137
συλλογισμός 8
επιχειρηματολογία 80
ισχυρισμός 5
επιχείρημα 1044

fax : 56
τηλεομοιοτυπία 3
φαξ 53

conference : 2518
συνδιάσκεψη 65
συνέδριο 69
διάσκεψη 2384

association : 834
συμμετοχή 11
σύνδεση 287
σύλλογος 48
συνεταιρίζεσθαι 41
οργάνωση 118
ένωση 329

paper : 387
βίβλος 72
έγγραφο 289
βιβλίο 26

adoption : 1066
έκδοση 52
υιοθέτηση 262
θέσπιση 57
έγκριση 665
λήψη 30

inventory : 38
απολογισμός 2
απογραφή 21
κατάλογος 15

assistance : 1544
βοήθεια 1004
ενίσχυση 301
συνδρομή 204
αρωγή 35

establishment : 795
ίδρυση 108
δημιουργία 217
κέντρο 12
καθιέρωση 107
εγκατάσταση 133
ίδρυμα 58
θέσπιση 160

attachment : 13
προσήλωση 13

variation : 24
τροποποίηση 1
μεταβολή 4
διακύμανση 19

settlement : 236
συμβιβασμός 6
οικισμός 40
διευθέτηση 116
διακανονισμός 74

manager : 77
διευθυντής 26
στέλεχος 12
διαχειριστής 39

preservation : 152
διατήρηση 118
διάσωση 1
διαφύλαξη 33

charge : 261
φόρος 51
τιμολόγηση 6
επιβάρυνση 56
τέλος 148

manufacture : 149
κατασκευή 44
παραγωγή 67
παρασκευή 38

nationality : 237
ιθαγένεια 26
υπηκοότητα 40
εθνικότητα 171

diversity : 581
ποικιλία 95
πολυμορφία 263
ποικιλότητα 36
ποικιλομορφία 187

bill : **98**
τιμολόγιο 5
νομοσχέδιο 34
λογαριασμός 59

replacement : **91**
ανανέωση 2
αναπλήρωση 1
αντικατάσταση 88

consent : 193
συμφωνία 9
έγκριση 17
συναίνεση 74
συγκατάθεση 93

awareness : 400
συνείδηση 109
ευαισθητοποίηση 167
συνειδητοποίηση 124

model : 1619
πρότυπο 986
υπόδειγμα 111
μοντέλο 522

facility : 342
υποδομή 19
εγκατάσταση 202
διευκόλυνση 102
εξοπλισμός 19

guardian : 118
θεματοφύλακας 112
κηδεμών 6

record : 70
μητρώο 10
καταγραφή 18
αρχείο 35
φάκελος 7

admission : 26
είσοδος 18
εισαγωγή 2
εισδοχή 6

contribution : 1894
συμμετοχή 96
συμβολή 979
συνεισφορά 546
εισφορά 267
εισήγηση 6

accommodation : 34
φιλοξενία 4
στέγαση 19
κατάλυμα 11

finding : 166
πόρισμα 84
εκτίμηση 4
διαπίστωση 58
εύρημα 20

momentum : 70
ταχύτητα 16
ορμή 15
ώθηση 39

duty : 1405
φόρος 268
υποχρέωση 139
καθήκον 789
δασμός 209

identification : 149
ταυτότητα 22
εντοπισμός 23
προσδιορισμός 15
ταυτοποίηση 14
αναγνώριση 75

waste : 1573
απόρριμμα 117
απόβλητο 1433
λύμα 23

overview : 49
ανασκόπηση 5
επισκόπηση 17
εικών 27

adequacy : 34
επάρκεια 33
καταλληλότητα 1

formulation : 114
διατύπωση 79
διαμόρφωση 35

care : 468
περίθαλψη 252
φροντίδα 120
μέριμνα 96

training : 1719
επιμόρφωση 76
κατάρτιση 1189
εκπαίδευση 454

definition : 1224
καθορισμός 178
ορισμός 965
προσδιορισμός 81

manufacturing : 68
κατασκευή 14
παραγωγή 37
παρασκευή 17

maintenance : 271
διατροφή 2
διατήρηση 184
συντήρηση 85

car : 865
αυτοκίνητο 781
αμάξι 3
όχημα 81

motivation : 95
κίνητρο 92
παρακίνηση 1
δραστηριοποίηση 1
ενθάρρυνση 1

disturbance : 14
διατάραξη 4
διαταραχή 10

response : 1560
αντίδραση 179
απάντηση 1214
ανταπόκριση 167

shortcoming : 322
κενό 39
έλλειψη 242
ανεπάρκεια 41

willingness : 207
προθυμία 128
επιθυμία 13
βούληση 66

restoration : 122
αποκατάσταση 108
αναστήλωση 2
ανάκτηση 8
αναζωογόνηση 0
συντήρηση 4

responsibility : 4644
μέριμνα 20
ευθύνη 4028
αρμοδιότητα 596

suggestion : 698
πρόταση 577
υπόδειξη 113
εισήγηση 8

removal : 70
απομάκρυνση 27
εξάλειψη 16
αφαίρεση 27

surveillance : 139
εποπτεία 23
επιτήρηση 36
επίβλεψη 21
παρακολούθηση 59

session : 1419
συνεδρίαση 377
συνεδρία 2
σύνοδος 1040

defect : 43
σφάλμα 4
ελάττωμα 32
ανωμαλία 7

equivalence : 36
αντιστοιχία 7
ισοτιμία 15
ισοδυναμία 14

assembly : 176
συνέρχεσθαι 4
συνέλευση 170
συνεταιρίζεσθαι 2

determination : 387
καθορισμός 18
διαπίστωση 2
προσδιορισμός 12
βούληση 123
αποφασιστικότητα 232

line : 1295
άξονας 41
κατεύθυνση 289
γραμμή 965

scope : 860
εύρος 70
έκταση 36
πεδίο 658
εμβέλεια 96

change : 3736
τροποποίηση 594
αλλαγή 2794
μεταβολή 348

sector : 6520
τομεύς 5518
κλάδος 377
τομέας 625

region : 6662
περιφέρεια 2232
περιοχή 4428
διαμέρισμα 2

difference : 1617
απόκλιση 58
διαφορετικότητα 6
διαφορά 1553

route : 301
οδός 75
διαδρομή 89
δρομολόγιο 37
δρόμος 100

disease : 1133
νόσος 253
νόσημα 10
ασθένεια 856
πάθηση 14

integration : 1902
ολοκλήρωση 1108
ένταξη 290
ενσωμάτωση 371
ενοποίηση 133

injury : 90
τραύμα 6
τραυματισμός 67
βλάβη 17

efficiency : 1001
απόδοση 139
αποδοτικότητα 144
αποτελεσματικότητα 718

arrangement : 646
ρύθμιση 394
καθεστώς 121
λεπτομέρεια 11
διευθέτηση 51
διακανονισμός 28
μηχανισμός 41

detection : 34
ανίχνευση 25
εξακρίβωση 1
διάγνωση 8

function : 411
καθήκον 103
αποστολή 17
λειτουργία 291

division : 188
διαχωρισμός 52
περιφέρεια 8
διαίρεση 61
κατανομή 67

range : 474
εύρος 29
φάσμα 196
ποικιλία 30
σειρά 219

channel : 167
οδός 21
κανάλι 90
δίαυλος 56

beach : 37
ακτή 13
παραλία 24

section : 532
τμήμα 356
ενότητα 1
μέρος 114
κεφάλαιο 61

registration : 235
καταγραφή 90
εγγραφή 33
καταχώρηση 71
καταχώριση 41

outcome : 912
πόρισμα 8
έκβαση 115
αποτέλεσμα 789

student : 406
σπουδαστής 121
φοιτητής 257
μαθητής 28

error : 586
σφάλμα 285
πλάνη 5
ελάττωμα 1
λάθος 295

film : 188
φιλμ 12
ταινία 176

education : 1458
κατάρτιση 98
παιδεία 183
εκπαίδευση 1158
αγωγή 19

movement : 1851
κυκλοφορία 1272
κίνηση 180
μετακίνηση 91
διακίνηση 308

relevance : 62
συνάφεια 8
σημασία 43
σπουδαιότητα 3
ενδιαφέρον 2
καταλληλότητα 6

release : 211
ελευθέρωση 61
απελευθέρωση 143
αποδέσμευση 7

prosperity : 522
ευημερία 503
ευμάρεια 19

promotion : 797
προβολή 10
προώθηση 631
προαγωγή 156

institution : 5894
όργανο 4575
φορέας 43
θεσμός 743
οργανισμός 178
ίδρυμα 355

amendment : 12658
τροποποίηση 1181
τροπολογία 11460
αναθεώρηση 17

street : 211
δρομάκι 1
δρόμος 210

designation : 72
διορισμός 2
ορισμός 4
ονομασία 66

rock : 14
πέτρα 5
βράχος 9

organisation : 2437
διοργάνωση 32
οργάνωση 1879
οργανισμός 526

coating : 1
επίστρωση 1

event : 1945
συμβάν 68
περιστατικό 27
εκδήλωση 131
γεγονός 1134
περίπτωση 585

completion : 194
συμπλήρωση 6
ολοκλήρωση 182
περάτωση 6

treatment : 1156
θεραπευτική 17
θεραπεία 166
επεξεργασία 118
αγωγή 17
μεταχείριση 698
αντιμετώπιση 140

belief : 193
πεποίθηση 153
πίστη 39
δοξασία 1

label : 374
σήμα 125
ετικέτα 140
επισήμανση 35
σήμανση 74

management : 2404
διαχείριση 2257
διοίκηση 93
διεύθυνση 54

circumstance : 1210
συνθήκη 719
περίσταση 167
περίπτωση 324

guidance : 125
προσανατολισμός 74
συμβουλή 7
καθοδήγηση 44

uncertainty : 448
ανασφάλεια 82
αβεβαιότητα 311
ασάφεια 30
αμφιβολία 25

administration : 1065
διαχείριση 285
διοίκηση 780

survey : 161
έρευνα 122
δημοσκόπηση 29
επισκόπηση 10

reference : 1359
μνεία 25
αναφορά 1244
παραπομπή 90

controversy : 41
διαμάχη 27
αντιπαράθεση 9
αντιδικία 5

specialisation : 19
ειδίκευση 1
εξειδίκευση 16
ειδικότητα 2

information : 7089
πληροφόρηση 1478
στοιχείο 389
πληροφορία 3879
ενημέρωση 1343

trade : 2905
εμπόριο 2304
επάγγελμα 24
συναλλαγή 577

attitude : 1102
συμπεριφορά 127
αντίληψη 28
στάση 947

player : 225
παράγοντας 147
φορέας 39
πρωταγωνιστής 34
συντελεστής 5

question : 11234
πρόβλημα 557
θέμα 2503
ζήτημα 2261
ερώτηση 3824
ερώτημα 2089

structure : 1787
κατασκευή 13
δομή 1358
διάρθρωση 416

communication : 3060
κοινοποίηση 12
ανακοίνωση 2119
επικοινωνία 929

D.3. Résultats du clustering pour les mots du lexique anglais-grec manuellement généré (non présentés dans le texte)

Mot ambigu	Clusters
communication	{ ενημέρωση, κοινοποίηση }
	{ διαβίβαση, ανταλλαγή }
	{ αναφορά }
	{ γνωστοποίηση, κοινοποίηση }
	{ επικοινωνία, ανακοίνωση }
	{ πληροφορία }
	{ κοινοποίηση, επικοινωνία }
paper	{ κείμενο }
	{ εργασία }
	{ δοκιμασία }
	{ χαρτί }
	{ έγγραφο, βιβλίο, βίβλος }
competence	{ ικανότητα, αρμοδιότητα }
	{ δυνατότητα }
	{ επάρκεια }
	{ δικαιοδοσία }
	{ δεξιότητα }
occupation	{ επάγγελμα }
	{ κατοχή, κατάληψη }
	{ ασχολία }
	{ απασχόληση }
	{ κατάληψη, δραστηριότητα }
preparation	{ παρασκεύασμα, παρασκευή, προετοιμασία }
	{ προετοιμασία, παρασκευή, επεξεργασία }
	{ προπαρασκευή }
	{ παρασκευή, κατάρτιση, επεξεργασία }
	{ επεξεργασία, κατάρτιση, εκπόνηση }
	{ παρασκεύασμα, σκεύασμα }
facility	{ ίδρυμα, υπηρεσία, δυνατότητα }
	{ μέσο, υποδομή }
	{ εξοπλισμός }
	{ δυνατότητα, υποδομή }
	{ διευκόλυνση, εγκατάσταση, υποδομή }

Mot ambigu	Clusters
treatment	{ διαχείριση }
	{ εξέταση }
	{ περίθαλψη, αντιμετώπιση, καθεστώς }
	{ κατεργασία }
	{ χορήγηση }
	{ θεραπεία, αγωγή }
	{ θεραπεία, μεταχείριση, επεξεργασία, περίθαλψη, αντιμετώπιση }
power	{ δικαιοδοσία }
	{ αρμοδιότητα, δύναμη, ισχύς, εξουσία }
	{ ενέργεια, ισχύς, δύναμη }
	{ ευχέρεια }
	{ εξουσία, δυνατότητα, αρμοδιότητα }
	{ καθήκον }

D.4. Résultats du clustering pour les mots du lexique anglais-grec automatiquement généré (non présentés dans le texte)

Mot ambigu	Clusters
department	{υπηρεσία, τμήμα, υπουργείο}
	{υπηρεσία, εξουσία}
	{εξουσία, νομός}
clothing	{ρούχο, ένδυση}
	{ιματισμός}
complaint	{καταγγελία, αναφορά}
	{ένσταση}
meal	{γεύμα, τροφή}
elimination	{αποβολή}
	{κατάργηση, εξάλειψη}
passage	{πάροδος, πέρασμα}
	{χωρίο}
opening	{απελευθέρωση, άνοιγμα}
	{διάνοιξη}
preparation	{παρασκεύασμα, παρασκευή, προετοιμασία}
	{εκπόνηση}
	{σκεύασμα, παρασκεύασμα}
conclusion	{συμπέρασμα, σύναψη}
	{πόρισμα}
issue	{πρόβλημα}
	{θέμα, ζήτημα}
	{έκδοση}
implication	{επίπτωση, συνέπεια}
	{επιπλοκή}
achievement	{επίδοση}
	{υλοποίηση, επίτευξη, επίτευγμα}
statement	{δήλωση, δικόγραφο}
	{υπόμνημα}
implementation	{εκτέλεση, υλοποίηση}
	{υλοποίηση, εφαρμογή}
foot	{πρόποδας, πόδι}
	{πόδας}
power	{δύναμη}
	{αρμοδιότητα, εξουσία}
	{ισχύς}
withdrawal	{αποχώρηση}
	{αφαίρεση, απόσυρση}
	{υπαναχώρηση}
	{απόσυρση, ανάκληση}
compliance	{τήρηση, συμμόρφωση}
	{σεβασμός}

period	{διάρκεια}
	{διάστημα, περίοδος, προθεσμία}
supplement	{συμπλήρωμα, συμπλήρωση}
	{προσθήκη}
renovation	{ανακαίνιση, ανανέωση}
	{αναπαλαίωση}
name	{όνομα, ονομασία}
	{επωνυμία}
interest	{ενδιαφέρον, ωφέλεια}
	{συμφέρον, ενδιαφέρον}
	{τόκος}
distribution	{κατανομή, διανομή}
	{διάδοση}
alcohol	{αλκοόλ, οινόπνευμα}
	{αλκοόλη}
introduction	{θέσπιση, καθιέρωση}
	{θέσπιση, κίνηση}
	{εισαγωγή, καθιέρωση}
consultation	{διαβούλευση, συνεννόηση}
	{διάλογος}
feast	{γιορτή, πανηγύρι}
	{εκδήλωση}
church	{εκκλησία, ναός}
	{εκκλησιά}
time	{προθεσμία, χρόνος, στιγμή}
	{χρόνος, φορά}
	{διάστημα, φορά}
area	{περιοχή, χώρος, τομεύς}
	{έκταση}
	{χώρος, πεδίο}
argument	{συλλογισμός, επιχειρηματολογία}
	{ισχυρισμός, επιχείρημα}
fax	{τηλεομοιοτυπία, φαξ}
conference	{συνέδριο, διάσκεψη}
	{συνδιάσκεψη}
association	{συμμετοχή, συνεταιρίζεσθαι}
	{σύλλογος, συμμετοχή, ένωση}
	{σύλλογος, οργάνωση}
	{συμμετοχή, ένωση, σύνδεση}
paper	{βιβλίο, βίβλος}
	{έγγραφο}
adoption	{έγκριση, θέσπιση, έκδοση}
	{λήψη}
	{υιοθέτηση, έγκριση, έκδοση}
inventory	{ευρετήριο, απολογισμός}
	{ευρετήριο, απογραφή}
	{κατάλογος, απογραφή}

establishment	{θέσπιση, καθιέρωση, δημιουργία, κέντρο}
	{ίδρυση}
	{ίδρυμα, καθιέρωση, δημιουργία, κέντρο}
	{ίδρυμα, εγκατάσταση, κέντρο}
variation	{διακύμανση}
	{μεταβολή, τροποποίηση}
assistance	{ενίσχυση, συνδρομή}
	{συνδρομή, βοήθεια}
	{αρωγή}
settlement	{διευθέτηση, διακανονισμός}
	{διακανονισμός, συμβιβασμός}
	{οικισμός}
manager	{διαχειριστής, διευθυντής}
	{στέλεχος, διευθυντής}
preservation	{διατήρηση, διαφύλαξη}
	{διάσωση}
charge	{τέλος, φόρος}
	{φόρος, επιβάρυνση}
	{τιμολόγηση}
manufacture	{παρασκευή, παραγωγή}
	{παραγωγή, κατασκευή}
nationality	{εθνικότητα, ιθαγένεια}
	{εθνικότητα, υπηκοότητα}
diversity	{ποικιλότητα}
	{πολυμορφία, ποικιλία, ποικιλομορφία}
bill	{τιμολόγιο, νομοσχέδιο}
	{λογαριασμός}
replacement	{ανανέωση, αναπλήρωση}
	{αντικατάσταση}
consent	{συναίνεση, συγκατάθεση}
	{συμφωνία}
	{έγκριση}
awareness	{συνειδητοποίηση ευαισθητοποίηση}
	{συνείδηση}
model	{πρότυπο, μοντέλο}
	{υπόδειγμα}
facility	{διευκόλυνση, υποδομή}
	{διευκόλυνση, εγκατάσταση}
	{εξοπλισμός}
record	{καταγραφή}
	{αρχείο, φάκελος}
	{μητρώο}
guardian	{επίτροπος}
	{θεματοφύλακας}
	{κηδεμόνας, κηδεμών}
admission	{εγγραφή}
	{εισαγωγή}

	{εισδοχή, είσοδος}
contribution	{συμμετοχή}
	{συνεισφορά, συμβολή}
	{εισφορά}
	{εισήγηση}
accommodation	{κατάλυμα, στέγαση}
finding	{εκτίμηση, διαπίστωση}
	{πόρισμα, εύρημα}
	{διαπίστωση, εύρημα}
duty	{καθήκον, υποχρέωση}
	{καθήκον, δασμός}
	{δασμός, φόρος}
identification	{εντοπισμός}
	{ταυτοποίηση}
	{ταυτότητα, προσδιορισμός, αναγνώριση}
waste	{λύμα, απόβλητο}
	{απόρριμμα}
overview	{επισκόπηση, εικών}
	{ανασκόπηση}
adequacy	{καταλληλότητα}
	{επάρκεια, ορθότητα}
formulation	{διαμόρφωση, χάραξη}
	{σύνθεση}
	{διατύπωση, διαμόρφωση}
care	{περίθαλψη, φροντίδα}
	{μέριμνα}
training	{κατάρτιση, εκπαίδευση}
	{επιμόρφωση, εκπαίδευση}
definition	{καθορισμός, ορισμός}
	{προσδιορισμός}
manufacturing	{παρασκευή, παραγωγή}
	{παραγωγή, κατασκευή}
maintenance	{συντήρηση, διατήρηση}
	{διατροφή}
motivation	{κίνητρο, δραστηριοποίηση}
	{παρακίνηση}
disturbance	{διαταραχή, διατάραξη}
response	{απάντηση, αντίδραση}
	{ανταπόκριση}
car	{αμάξι}
	{αυτοκίνητο, όχημα}
shortcoming	{ανεπάρκεια}
	{κενό, έλλειψη}
willingness	{βούληση}
	{επιθυμία, προθυμία}
responsibility	{ευθύνη, αρμοδιότητα}
	{μέριμνα}

suggestion	{υπόδειξη, πρόταση}
removal	{εξάλειψη}
	{αφαίρεση, απομάκρυνση}
surveillance	{εποπτεία}
	{παρακολούθηση, επιτήρηση, επίβλεψη}
session	{συνεδρίαση, σύνοδος}
	{συνεδρία}
defect	{σφάλμα}
	{ανωμαλία, ελάττωμα}
equivalence	{αντιστοιχία}
	{ισοδυναμία, ισοτιμία}
assembly	{συνέρχεσθαι, συνέλευση}
determination	{διαπίστωση, βούληση}
	{αποφασιστικότητα}
	{διαπίστωση, καθορισμός}
	{προσδιορισμός}
line	{κατεύθυνση, άξονας}
	{γραμμή}
scope	{πεδίο, εμβέλεια}
	{εμβέλεια, έκταση, εύρος}
change	{μεταβολή, αλλαγή}
	{τροποποίηση}
sector	{κλάδος, τομέας}
	{τομείς}
region	{περιοχή, περιφέρεια}
	{διαμέρισμα}
difference	{απόκλιση, διαφορά}
	{διαφορετικότητα}
route	{διαδρομή, οδός}
	{δρομολόγιο}
	{δρόμος}
disease	{πάθηση, νόσημα, νόσος}
	{νόσος, ασθένεια}
integration	{ενοποίηση}
	{ένταξη, ενσωμάτωση, ολοκλήρωση}
injury	{τραυματισμός, βλάβη}
	{τραύμα}
arrangement	{λεπτομέρεια, διευθέτηση, ρύθμιση}
	{ρύθμιση, μηχανισμός}
	{ρύθμιση, καθεστώς, λεπτομέρεια}
	{ρύθμιση, διακανονισμός, καθεστώς}
efficiency	{απόδοση}
	{αποδοτικότητα, αποτελεσματικότητα}
detection	{ανίχνευση, διάγνωση}
	{εξακρίβωση}
function	{λειτουργία, καθήκον}
	{αποστολή}

division	{διαίρεση, κατανομή}
	{κατανομή, διαχωρισμός}
	{περιφέρεια}
range	{σειρά, φάσμα}
	{φάσμα, ποικιλία}
	{εύρος}
section	{ενότητα, μέρος}
	{μέρος, τμήμα, κεφάλαιο}
beach	{ακτή}
	{παραλία, αμμουδιά}
registration	{εγγραφή, καταχώρηση}
	{εγγραφή, καταχώριση}
	{καταγραφή}
outcome	{έκβαση, αποτέλεσμα}
	{πόρισμα}
student	{σπουδαστής, φοιτητής}
	{μαθητής, σπουδαστής}
error	{σφάλμα, πλάνη}
	{λάθος}
	{ελάττωμα}
education	{εκπαίδευση}
	{κατάρτιση, παιδεία, αγωγή}
movement	{κυκλοφορία}
	{μετακίνηση, κίνηση, διακίνηση}
relevance	{σπουδαιότητα}
	{καταλληλότητα, συνάφεια}
	{ενδιαφέρον}
	{συνάφεια, σημασία}
release	{αποδέσμευση, ελευθέρωση}
	{απελευθέρωση}
prosperity	{ευημερία, ακμή}
	{ευμάρεια}
promotion	{προώθηση, προαγωγή}
	{προβολή}
institution	{θεσμός, οργανισμός}
	{όργανο, ίδρυμα}
	{οργανισμός, φορέας}
amendment	{τροποποίηση, τροπολογία}
	{αναθεώρηση}
street	{δρομάκι, σοκάκι}
	{δρομάκι, δρόμος}
designation	{ονομασία, χαρακτηρισμός}
	{διορισμός, χαρακτηρισμός}
	{ορισμός}
rock	{πέτρωμα, βράχος}
	{πέτρα}
organisation	{οργανισμός, οργάνωση}

	{διοργάνωση}
event	{γεγονός, περίπτωση, εκδήλωση}
	{συμβάν}
	{γεγονός, περιστατικό}
completion	{ολοκλήρωση, συμπλήρωση}
	{περάτωση}
treatment	{θεραπεία, αγωγή}
	{επεξεργασία, μεταχείριση, θεραπεία}
	{θεραπευτική, αντιμετώπιση}
belief	{δοξασία}
	{πεποίθηση, πίστη}
label	{επισήμανση, ετικέτα}
	{σήμα}
	{επισήμανση, σήμανση}
management	{διεύθυνση, διοίκηση}
	{διοίκηση, διαχείριση}
circumstance	{συνθήκη, περίσταση}
	{περίπτωση}
guidance	{προσανατολισμός, καθοδήγηση}
	{συμβουλή}
administration	{διοίκηση, διαχείριση}
	{χορήγηση, διοίκηση}
coating	{επίχριση}
	{επίχρισμα, επίστρωση}
survey	{επισκόπηση}
	{δημοσκόπηση, έρευνα}
reference	{παραπομπή, αναφορά}
	{παραπομπή, μνεία}
specialisation	{ειδικότητα, ειδίκευση}
	{εξειδίκευση}
information	{στοιχείο, ενημέρωση}
	{πληροφόρηση, ενημέρωση}
	{πληροφορία}
trade	{συναλλαγή, εμπόριο}
	{επάγγελμα}
attitude	{συμπεριφορά, στάση}
	{αντίληψη}
player	{πρωταγωνιστής}
	{φορέας, παράγοντας, συντελεστής}
question	{ζήτημα, ερώτημα, θέμα}
	{πρόβλημα}
	{θέμα, ερώτηση}
structure	{διάρθρωση, δομή}
	{κατασκευή}
communication	{επικοινωνία, ανακοίνωση}
	{κοινοποίηση}
penetration	{διάτρηση}

	{διείσδυση}
	{εισχώρηση}
controversy	{αντιδικία}
	{αντιπαράθεση}
	{διαμάχη}
uncertainty	{αμφιβολία}
	{αβεβαιότητα}
	{ασάφεια}
	{ανασφάλεια}
film	{μεμβράνη}
	{επίστρωση}
	{ταινία}
	{φιλμ}
channel	{κανάλι}
	{δίαυλος}
	{οδός}
momentum	{ταχύτητα}
	{ορμή}
	{ώθηση}
attachment	{κατάσχεση}
	{προσάρτημα}
	{προσήλωση}

Aucune relation de similarité sémantique n'a été repérée entre les équivalents des mots suivants : *controversy*, *uncertainty*, *film*, *channel*, *momentum*, *attachment*, *penetration*, *restoration*, *multitude*, *reorganization*. Ce fait explique que chacun des équivalents de ces mots est inclus dans un cluster tout seul, aucun regroupement n'ayant été effectué.

Annexe E

E.1. Table de translittération des caractères grecs en caractères latins

ISO[1] 843 : 1997 TR[(2.0)]

Α	α	a
	αι	ai
	άι	ái
	αϊ	aï
	αυ	av, af, ay[(2.1)]
Β	β	v
Γ	γ	g
	γγ	ng
	γκ	gk
	γξ	nx
	γχ	nch
Δ	δ	d
Ε	ε	e
	ει	ei
	έι	éi
	εϊ	eï
	ευ	ev, ef, ey[(2.1)]
Ζ	ζ	z
Η	η	i
	ηυ	iv, if, iy[(2.1)]
Θ	θ	th
Ι	ι	i
Κ	κ	k
Λ	λ	l
Μ	μ	m
	μπ	b, mp[(2.2)]
Ν	ν	n
	ντ	nt
Ξ	ξ	x
Ο	ο	o

[1] International Standards Organisation.

	οι	oi
	όi	όi
	οϊ	οϊ
	ου	ou, oy[2.3]
Π	π	p
Ρ	ϱ	r
Σ	σ, ς	s
Τ	τ	t
Υ	υ	y
	υι	yi
Φ	ϕ	f
Χ	χ	ch
Ψ	ψ	ps
Ω	ω	o

E2. Jeu d'étiquettes utilisé pour l'étiquetage morphosyntaxique du grec

Etiquette	Partie du discours
ABBR	Abbreviation
AdXxBa	Adverb, Basic form
AdXxCp	Adverb, Comparative form
AdXxSu	Adverb, Superlative form
AjBa	Adjective, Basic form
AjCp	Adjective, Comparative form
AjSu	Adjective, Superlative form
AsPpPa	Prepart Preposition
AsPpSp	Simple Preposition
AtDf	Definite Article
AtId	Indefinite Article
CjCo	Coordinate Conjunction
CjSb	Subordinate Conjunction
DATE	Date
DIG	Digit
ENUM	Enumeration
Ij	Interjection
INIT	Initials
NBABBR	Non-breaking Abbreviation
NmAn	Analog Numeral
NmCd	Cardinal Numeral
NmCt	Collective Numeral
NmMl	Multiplicative Numeral
NmOd	Ordinal Numeral
NoCm	Noun Common

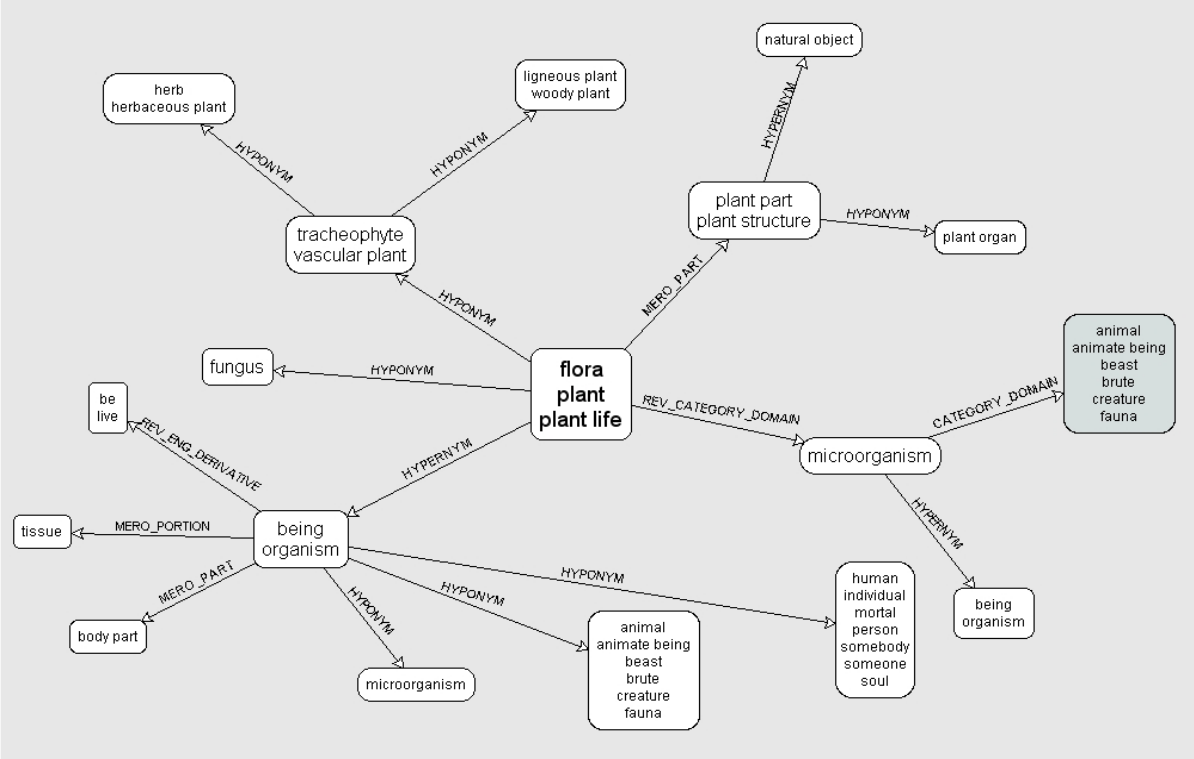
NoPr	Noun Proper
PnDm	Demonstrative Pronoun
PnId	Indefinite Pronoun
PnIr	Interrogative Pronoun
PnPe	Personal Pronoun
PnPo	Possessive Pronoun
PnRe	Relative Pronoun
PnRi	Relative Indefinite Pronoun
PtFu	Future Particle
PtNg	Negative Particle
PtOt	Other Particles
PtSj	Subjunctive Particle
Pu	Punctuation
PUNCT	Punctuation
RgAbXx	Abbreviation
RgAnXx	Acronym
RgFwOr	Original Foreign Word
RgFwTr	Transliterated Foreign Word
RgSyXx	Symbol
TE	Tagging Error
VbIs	Impersonal Verb
VbMn	Main Verb

Annexe F

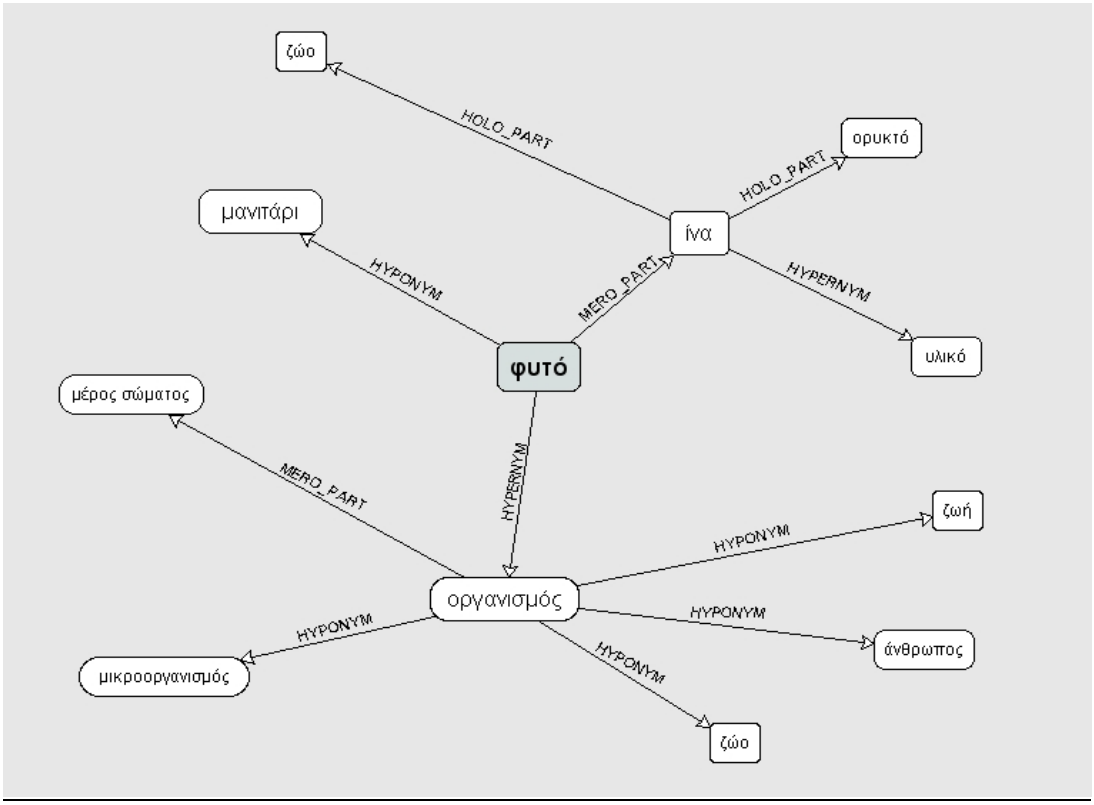
F1. Représentation des sens de *plant* dans BalkaNet

Sens 1 : {**flora, plant, plant life**} : a living organism lacking the power of locomotion (*organisme vivant dépourvu de capacité de mouvement*)

AN :

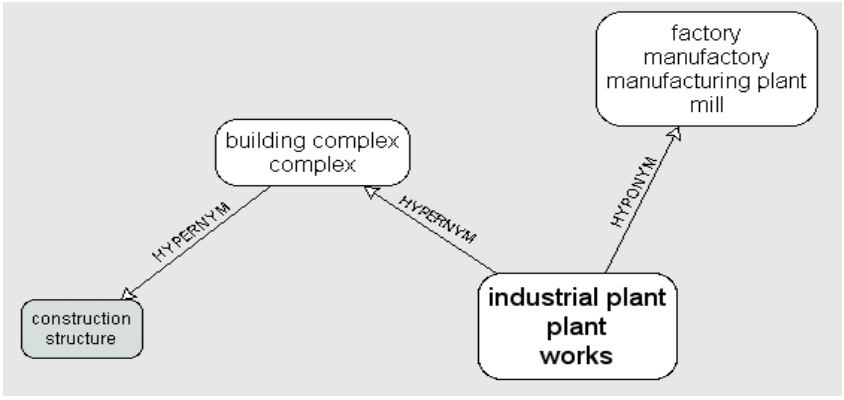


GR

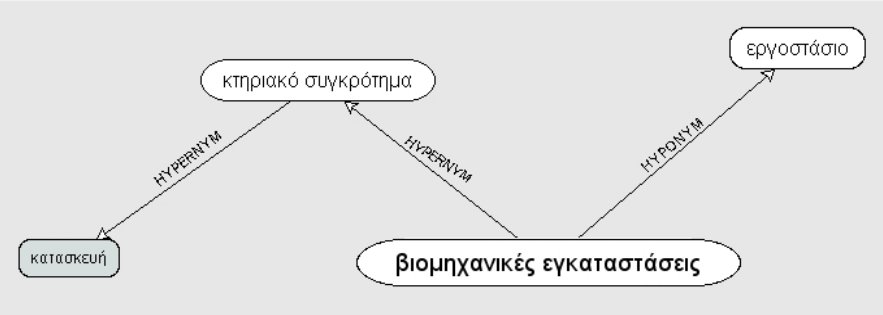


Sens 2 : {**industrial plant, plant, works**} : buildings for carrying on industrial labor (*bâtiments consacrés aux activités industrielles*)

AN :



GR :



Sens 3 : {**plant**} : something planted secretly for discovery by another (*ce qui est caché secrètement (par quelqu'un) pour être découvert par quelqu'un d'autre*)

Sens 4 : {**plant**} : an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience (*acteur situé dans le public et dont le jeu a été travaillé mais paraît improvisé au public*)

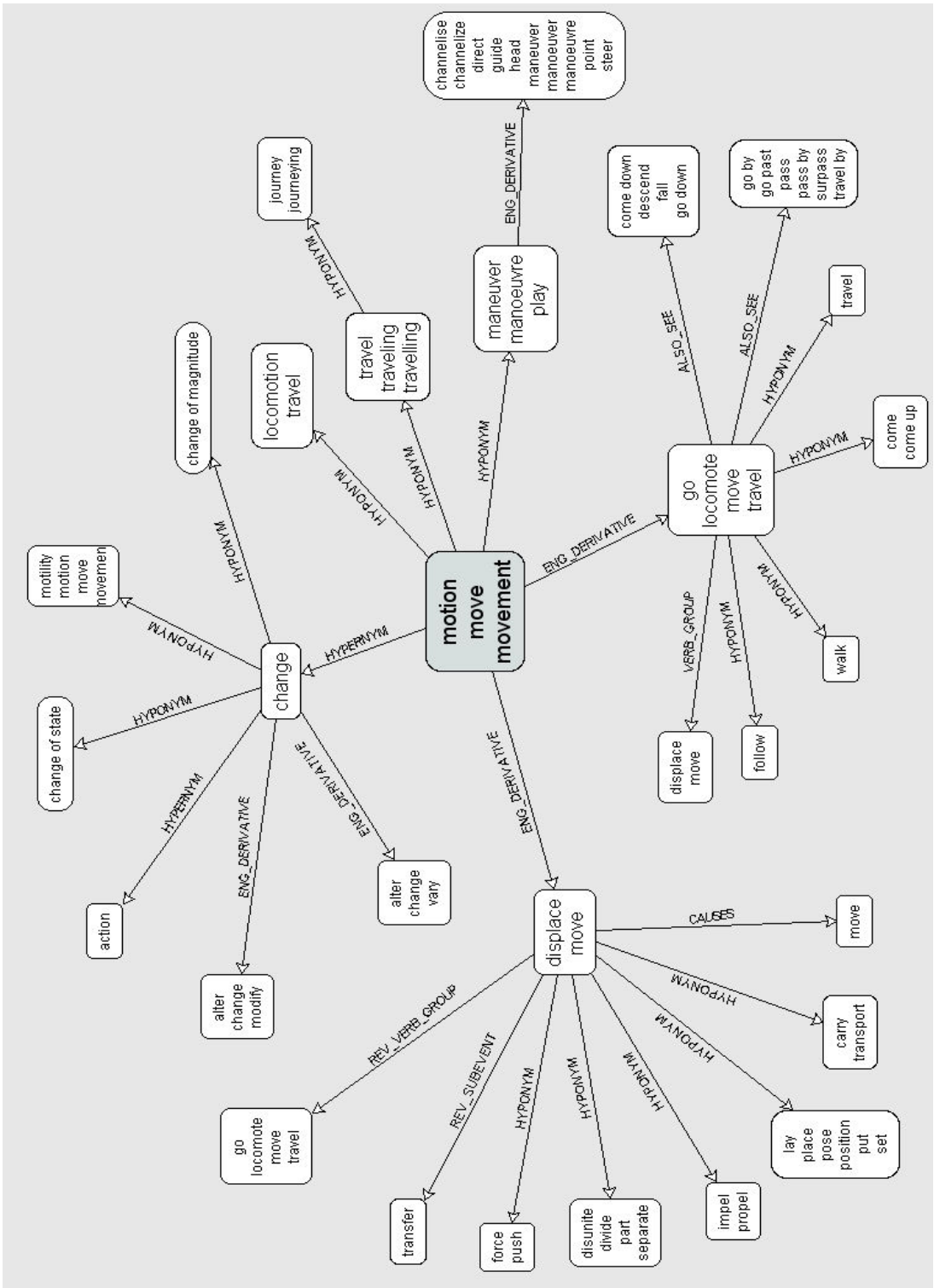
***Les sens 3 et 4 sont décrits au sein du réseau AN par un seul synset, contenant le mot* plant.**

Sens 3 : **GR**

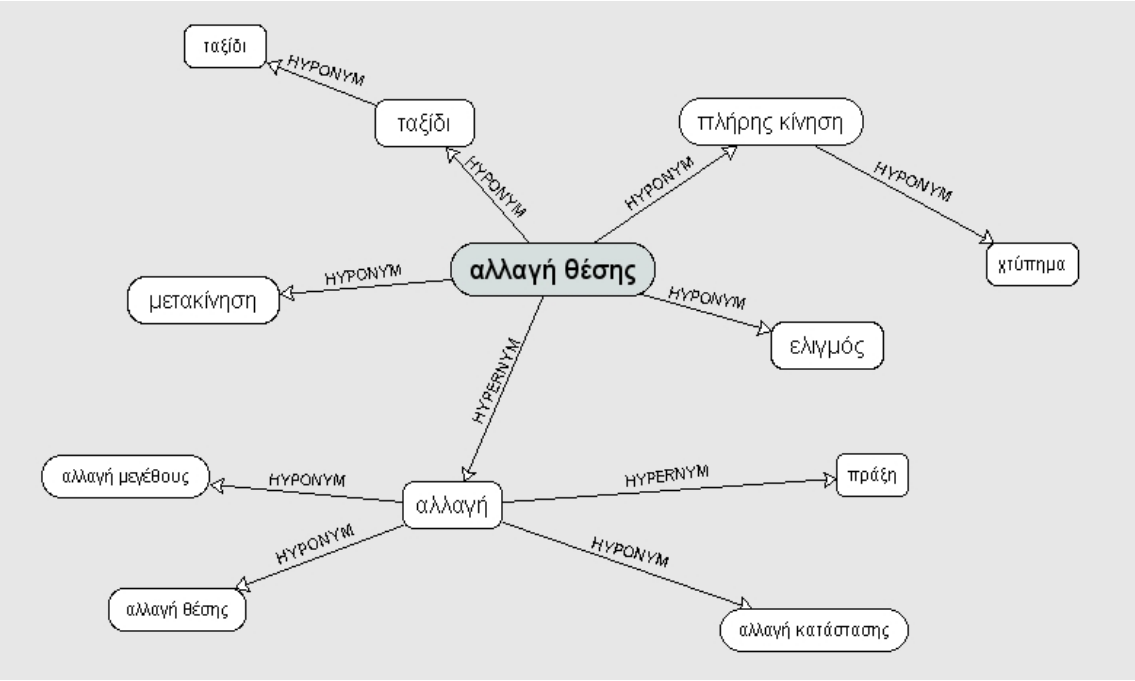
Sens 4 : **GR**

F2. Représentation des sens de *movement* dans BalkaNet

Sens 1 : {**motion, move, movement**} : the act of changing location from one place to another (*action de changer de lieu, d'aller d'un endroit à un autre*)
EN

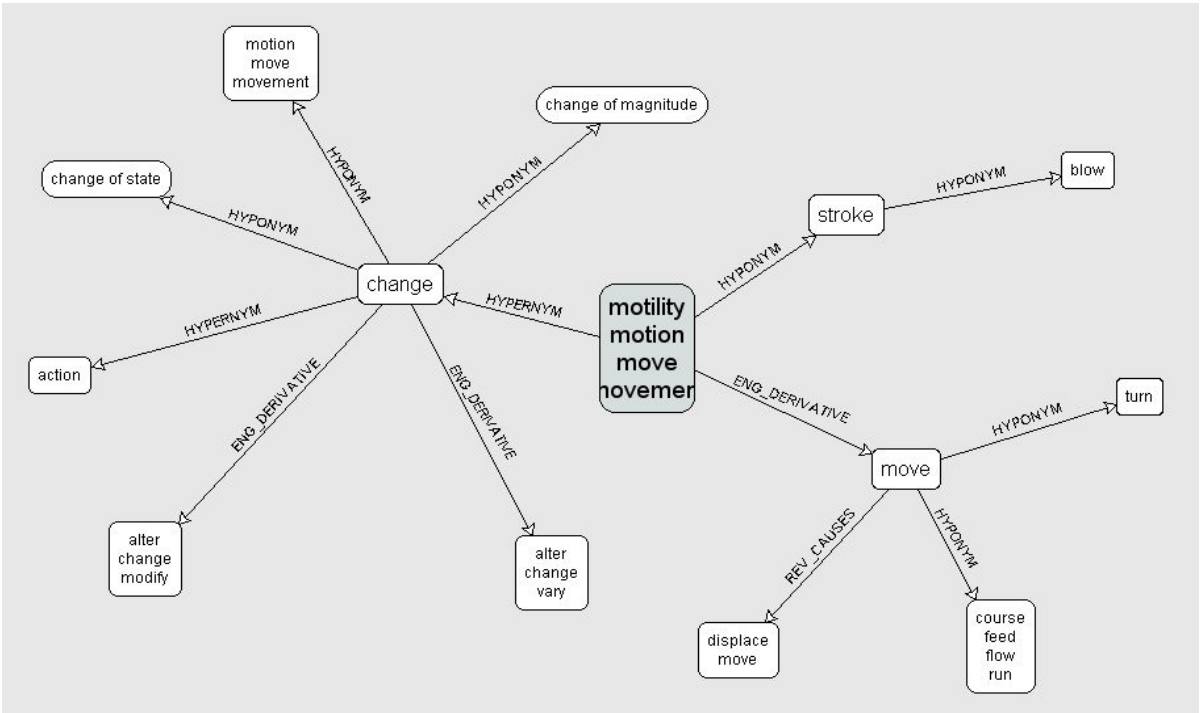


(Sens1) GR

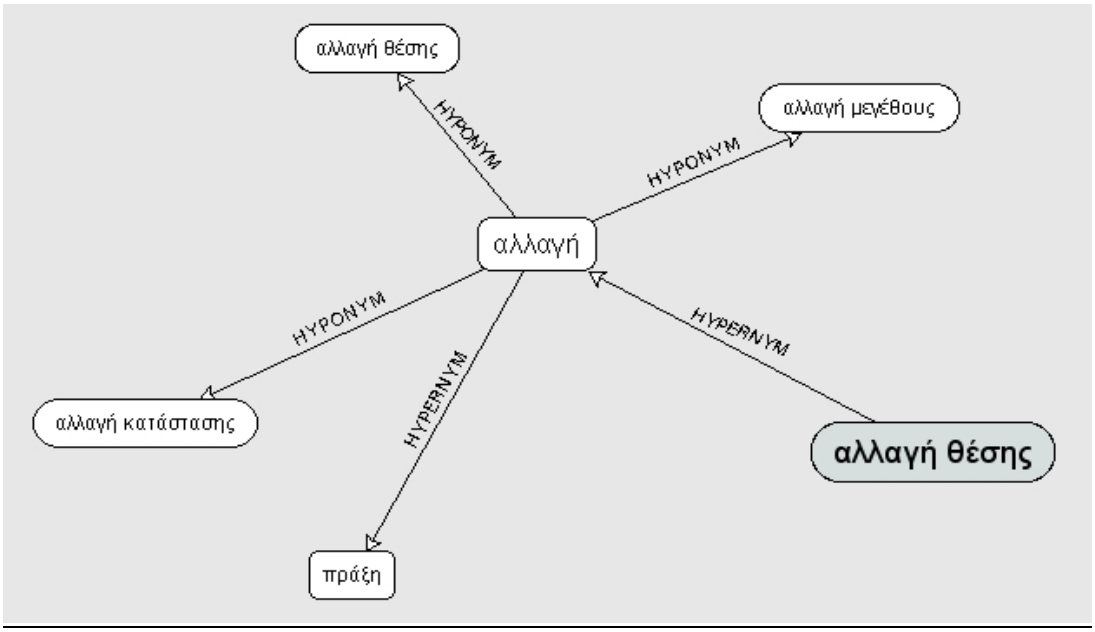


Sens 2 : {**motility, motion, move, movement**} : a change of position that does not entail a change of location (*changement de position qui n'entraîne pas un changement de lieu*)

EN

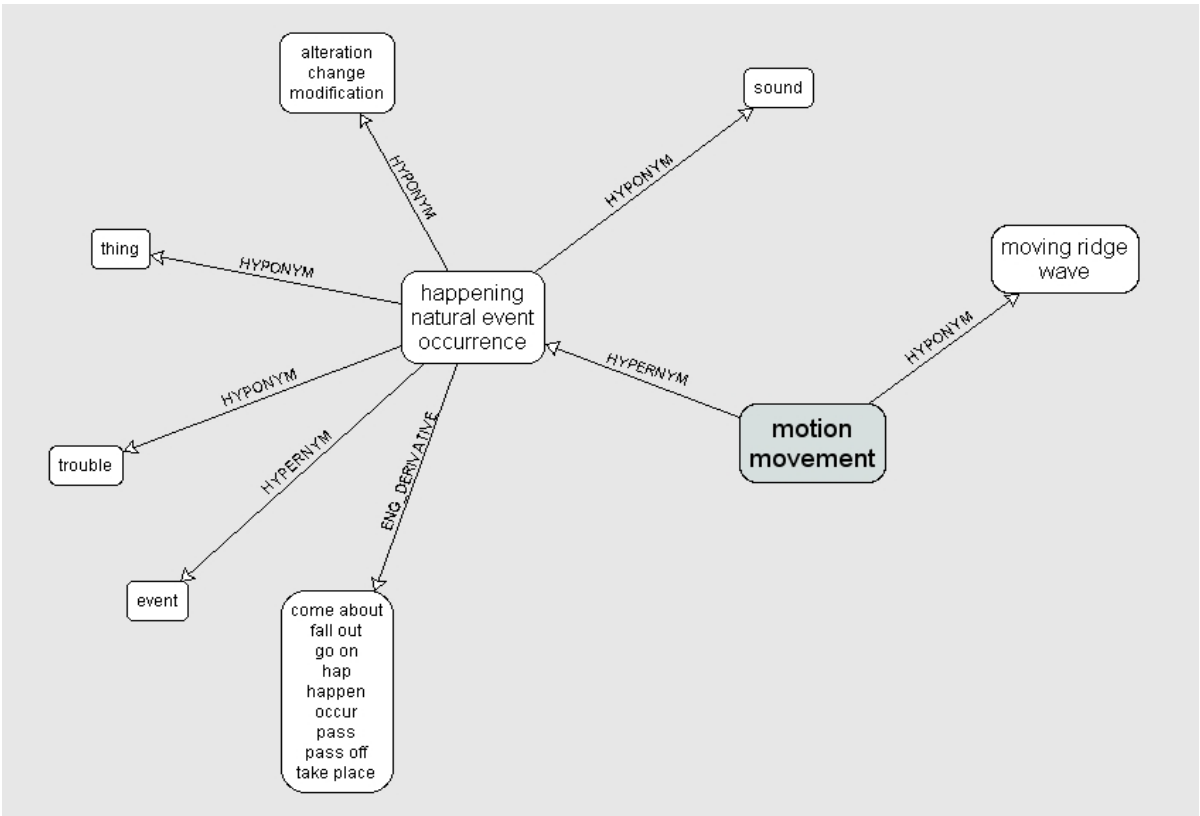


(sens 2) GR

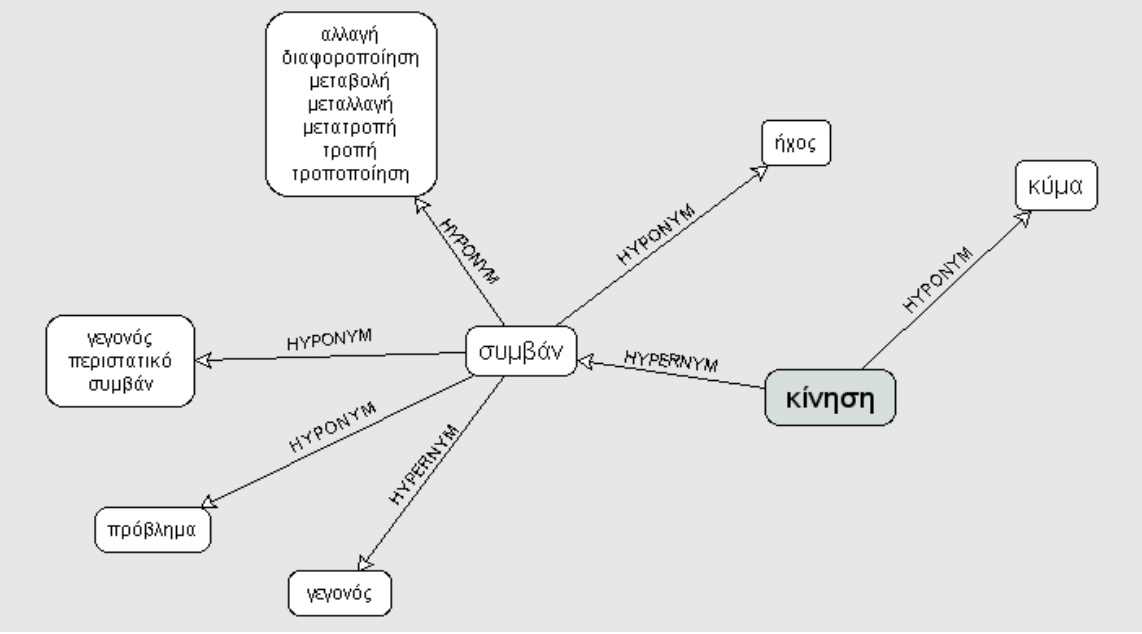


Sens 3 : {**motion, movement**} : a natural event that involves a change in the position or location of something (*événement naturel qui implique un changement de position ou de lieu de quelque chose*)

EN

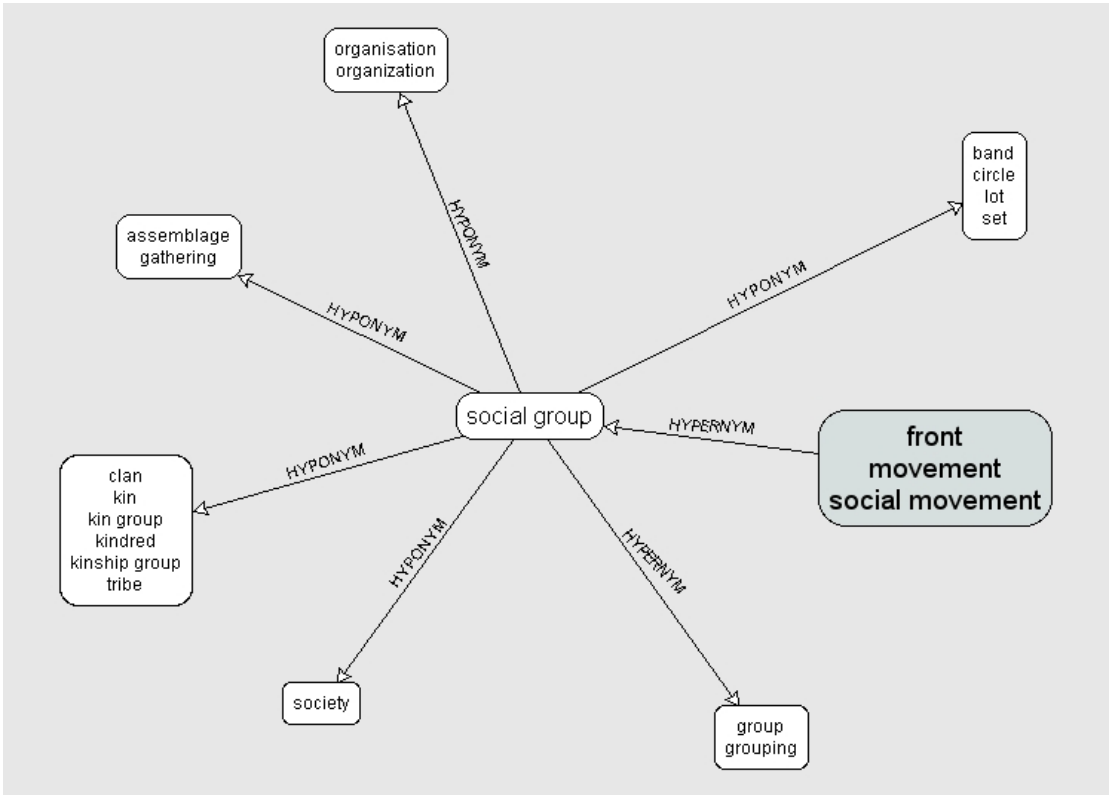


(sens 3) GR

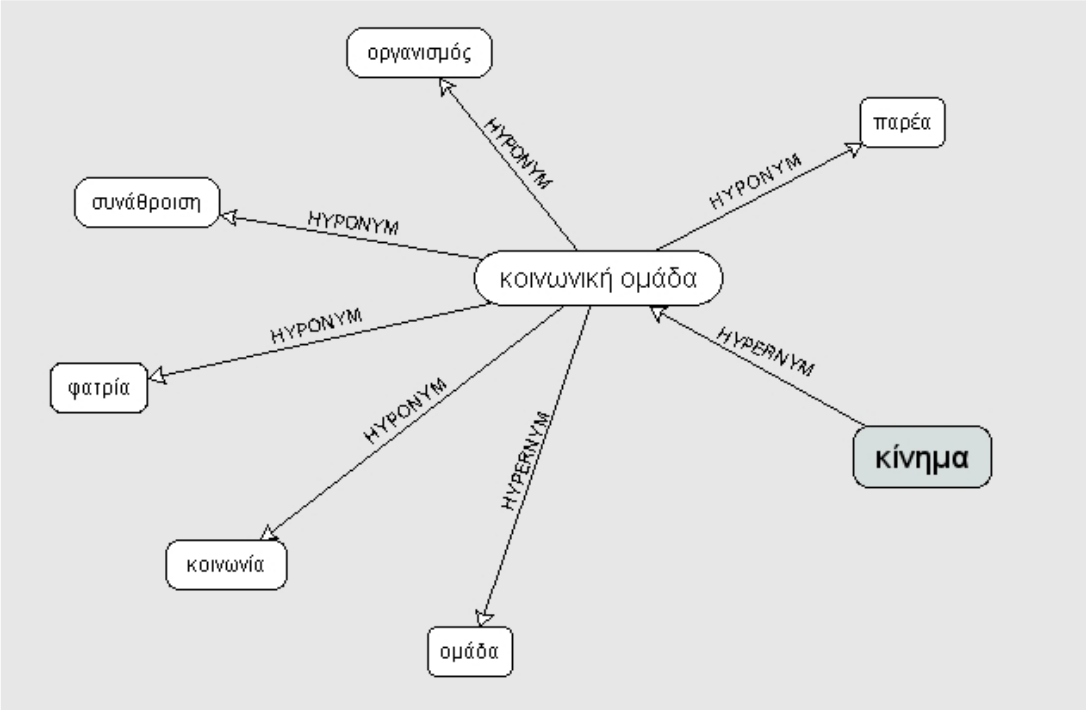


Sens 4 : {**front, movement, social movement**} : a group of people with a common ideology who try together to achieve certain general goals (*groupe de personnes ayant une idéologie commune, qui essaient ensemble d'atteindre certains buts généraux*)

EN



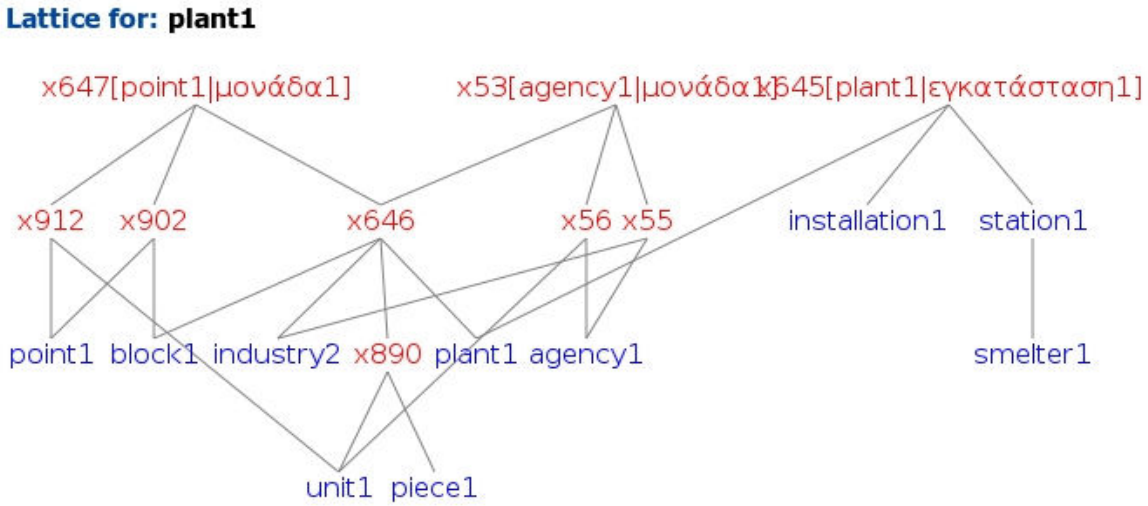
(sens 4) GR


οργανισμός
HYPONYM
παρέα
HYPONYM
συνάθροιση
HYPONYM
κοινωνική ομάδα
HYPONYM
φατρία
HYPONYM
HYPERNYM
κίνημα
HYPONYM
κοινωνία
HYPERNYM
ομάδα

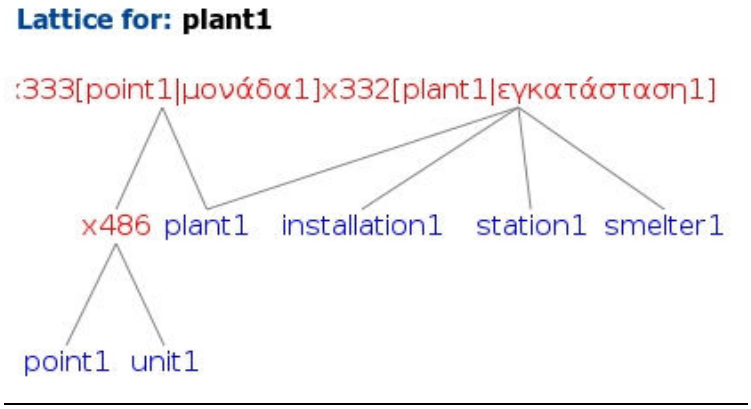
F3. Semi-treillis construits pour les mots *plant* et *movement*

plant

(*prise en compte des EQVs de traduction qui apparaissent une seule fois* (hapax))



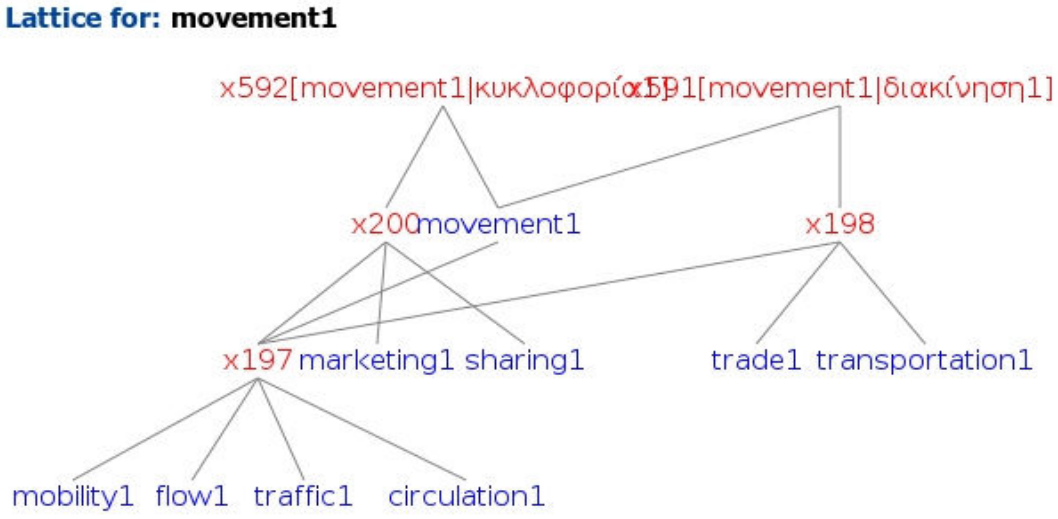
(*non prise en compte des EQVs de traduction qui apparaissent une seule fois* (hapax))



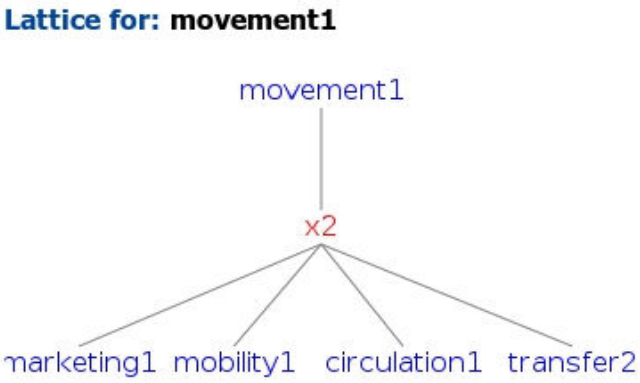
---

movement

(*prise en compte des* hapax)



(*non prise en compte des* hapax)



F4. Thesaurus généré à partir du lexique bilingue manuellement construit

Nous présentons ici le thesaurus construit par la méthode des Miroirs Sémantiques à partir de notre lexique bilingue manuellement généré lors des trois inversions de la direction de traduction. En raison de limitations d'espace, nous excluons les entrées du thesaurus et les sens lexicaux étant illustrés uniquement par des traductions (c'est-à-dire pour lesquels des relations sémantiques n'ont pas été repérées). Nous avons également exclu les entrées contenant un seul sous-sens, aussi bien que celles construites pour les traductions repérées lors de la troisième inversion de la direction de traduction (appelées *starred entries*). Les informations disponibles concernant les traductions, repérées lors du dernier tour, sont incomplètes, étant donné que leurs relations de traduction n'ont pas été extraites à partir du corpus.

Néanmoins, il faut noter que le thesaurus complet, comprenant les types d'entrées que nous avons exclus ici, est consultable sur le site Web des Miroirs (http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html), qui contient la totalité des résultats obtenus par l'application de la méthode des Miroirs sur nos données.

Mirrors-Web : Thesaurus Listing

Word Base: **enggreekmerged** extended (from 0, totally 248 entries, 0 starred entries)
Synset Limit: **automatic**
Overlap Threshold: **0.05**

ability
Subsense (i)
(*Translation:* δεξιότητα.)
Synonyms: know-how.
Subsense (ii)
(*Translation:* ικανότητα, δυνατότητα.)
Synonyms: capability‹1›, capacity‹1›, competence, discretion‹1›, facility‹1›, power‹1›, qualification, resource‹1›, right‹1›, skill‹1›.
Related words: aptitude‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.
act
Hyperonyms: development‹1›.
Subsense (i)
(*Translation:* περισταστικό.)
Synonyms: accident.
Subsense (ii)
(*Translation:* μέτρο, ενέργεια, διαδικασία.)
Synonyms: action, extent, instrument, legislation, means, measure‹1›, remedy‹1›, rule‹1›, standard, step‹1›.
Related words: attempt, course‹1›, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, force‹1›, goal‹1›, intervention, power‹1›, project.
Subsense (iii)
(*Translation:* πράξη.)
Synonyms: document.
Related words: concentration‹1›, order‹1›, practice.
activity
Sense 1
(*Translation:* εργασία, διαδικασία.)
Hyperonyms: development‹1›.
Hyponyms: action, assignment‹1›, attempt, discussion‹1›, document, effort‹1›, employment, energy‹1›, exercise‹1›, intervention, job, labour, motivating, occupation‹1›, operating, operation, paper, posting, procedure‹1›, proceedings, project, report, resolution, service, study, task, work, working.
Sense 2
(*Translation:* ραδιενέργεια.)
Synonyms: radiation.
administering
Subsense (i)
(*Translation:* χορήγηση, παροχή, λήψη, απονομή.)

Synonyms: administration‹1›, allocation‹1›, attribution, awarding‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, paying‹1›, provision‹1›.
Related words: award‹1›, benefit‹1›, dosing‹2›, intake‹3›, issue‹2›, obligation, obtaining‹1›, performance‹1›, providing‹1›, receipt, receiving, reception‹1›, sourcing‹1›, supply‹1›, taking‹2›, therapy, treatment‹1›, use.
Subsense (ii)
(*Translation:* διοίκηση, διαχείριση.)
Synonyms: governance, government, management‹1›.
Related words: command‹1›, conduct‹1›, handling, processing‹1›, running.

administration
Subsense (i)
(*Translation:* χορήγηση, λήψη.)
Synonyms: administering, grant‹1›, intake‹3›, obtaining‹1›, provision‹1›, receipt, receiving, reception‹1›, sourcing‹1›, taking‹2›.
Related words: allocation‹1›, attribution, award‹1›, awarding‹1›, delivery‹1›, dosing‹2›, giving‹1›, granting‹1›, issue‹2›, issuing, paying‹1›, therapy, treatment‹1›, use.
Subsense (ii)
(*Translation:* υπηρεσία, οργανισμός.)
Synonyms: agency, body‹1›, office‹1›, service.
Related words: authority‹2›, centre, department, duty‹1›, employment, entity, facility‹1›, force‹1›, input‹2›, institution, organisation‹1›, platform, point, sector, unit, work.

aid
Subsense (i)
(*Translation:* ενίσχυση, συνδρομή.)
Synonyms: assistance‹1›, assisting, contribution, donation, help, helping, input‹3›, support.
Related words: enhancement, enhancing‹1›, generating, grant‹1›, improvement, improving‹1›, increase‹1›, increasing, strengthening.
Subsense (ii)
(*Translation:* μέσο.)
Synonyms: instrument, means, middle, possibility, remedy‹1›, resort, resource‹1›, tool‹1›, way.

allocation
Subsense (i)
(*Translation:* χορήγηση, παροχή.)
Synonyms: administering, awarding‹1›, benefit‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, obligation, paying‹1›, performance‹1›, providing‹1›, provision‹1›, supply‹1›.
Related words: administration‹1›, attribution, award‹1›, dosing‹2›, issue‹2›, therapy, treatment‹1›, use.
Subsense (ii)
(*Translation:* θέση.)
Synonyms: role‹1›.
Related words: attention, job, location‹1›, office‹1›, opinion, place, placing, position, post‹1›, release‹1›, seat, settlement‹1›, site, situation‹1›, spot, status‹1›.

application
(*Translation:* υλοποίηση, λειτουργία, εφαρμογή.)
Hyperonyms: arrangement‹1›.
Synonyms: achievement, achieving, attaining, attainment‹1›, birth, completion‹1›, conduct‹1›, execution‹1›, implementation, implementing, performing, preparation‹1›, pursuance, pursuing, realisation, task.
Related words: effect‹1›, function‹1›, functioning, governance, habit‹1›, management‹1›, operating, performance‹1›, process, purpose, running, service, significance, status‹1›, working.

applying
Hyperonyms: arrangement‹1›.
Subsense (i)
(*Translation:* χρήση, εφαρμογή.)
Synonyms: use.
Related words: consumption‹2›, dosing‹2›, exploitation‹1›, management‹1›, operating, purpose, recourse‹1›, sourcing‹1›, taking‹2›, treatment‹1›, usage‹2›, using, work, working.
Subsense (ii)
(*Translation:* αίτηση.)
Synonyms: claim‹1›.
Related words: appeal, complaint, demand, petition, request‹1›.

aptitude
Subsense (i)
(*Translation:* δεξιότητα.)
Synonyms: know-how.
Subsense (ii)
(*Translation:* ικανότητα, αρμοδιότητα.)
Synonyms: competence, power‹1›, qualification.
Related words: ability, assignment‹1›, capability‹1›, capacity‹1›, discretion‹1›, efficiency‹1›, expertise‹2›, function‹1›, jurisdiction‹1›, mandate, performance‹1›, property‹2›, quality, remit, resource‹1›, responsibility‹1›, right‹1›, scope‹1›, skill‹1›, status‹1›.

arrangement
(*Translation:* εφαρμογή.)
Hyperonyms: development‹1›.
Hyponyms: appliance, application‹1›, applying‹1›, effect‹1›, enforcement, execution‹1›, harnessing, implementation, imposition, introducing, introduction‹1›, launching, operating, operation, practice, pursuance, resolution, spread, use, using, utilisation.

assessment
(*Translation:* μελέτη, έλεγχος, εξέταση.)
Hyperonyms: event‹1›, consideration‹1›.
Synonyms: analysis, assay, consultation‹1›, debate, discussing, discussion‹1›, document, exercise‹1›, investigation‹1›, project, reading‹1›, reflection‹2›, report, study, survey‹1›, trial.

assignment
Hyperonyms: activity‹1›.
Subsense (i)
(*Translation:* έργο.)
Synonyms: achievement, analysis, contribution, record‹1›.
Subsense (ii)
(*Translation:* αρμοδιότητα.)
Synonyms: power‹1›.
Related words: aptitude‹1›, competence, expertise‹2›, function‹1›, jurisdiction‹1›, mandate, remit, responsibility‹1›, scope‹1›.

assistance
(*Translation:* ενίσχυση.)
Synonyms: aid‹1›, enhancement, enhancing‹1›, generating, grant‹1›, improvement, improving‹1›, increase‹1›, increasing, strengthening, support.

attention
Subsense (i)
(*Translation:* σημασία, ενδιαφέρον.)

Synonyms: importance, interest‹1›, scope‹1›, significance, usefulness, value.
Related words: extent, implications, meaning, relevance‹1›, sense‹1›, stress‹1›.
Subsense (ii)
(*Translation:* θέση.)
Synonyms: role‹1›.
Related words: allocation‹1›, job, location‹1›, office‹1›, opinion, place, placing, position, post‹1›, release‹1›, seat, settlement‹1›, site, situation‹1›, spot, status‹1›.

authorisation
(*Translation:* άδεια.)
Synonyms: authorization.
Hyponyms: approval, authority‹1›, leave‹1›, licensing, permission.

authority
Sense 1
(*Translation:* άδεια.)
Hyperonyms: authorisation‹1›.
Sense 2
(*Translation:* αρχή, υπηρεσία.)
Synonyms: agency, association, beginning, body‹1›, commencement, competence, concept‹1›, entity, policy‹1›, principle‹1›, right‹1›, rule‹1›, service, start, term, value.
Related words: administration‹1›, centre, department, duty‹1›, employment, facility‹1›, force‹1›, input‹2›, office‹1›, platform, point, sector, unit, work.

award
(*Translation:* απονομή, ανάθεση, χορήγηση.)
Synonyms: attribution, awarding‹1›, giving‹1›, provision‹1›.
Related words: administering, administration‹1›, allocation‹1›, attachment‹1›, conferring, contracting, delegation‹1›, delivery‹1›, dosing‹2›, grant‹1›, granting‹1›, issue‹2›, issuing, paying‹1›, therapy, treatment‹1›, use.

awarding
(*Translation:* ανάθεση, παροχή, χορήγηση.)
Synonyms: administering, allocation‹1›, attachment‹1›, attribution, award‹1›, conferring, contracting, delegation‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, paying‹1›, provision‹1›.
Related words: administration‹1›, benefit‹1›, dosing‹2›, issue‹2›, obligation, performance‹1›, providing‹1›, supply‹1›, therapy, treatment‹1›, use.

burden
Subsense (i)
(*Translation:* βαρύτητα, βάρος.)
Synonyms: barrier, value, weight.
Related words: importance.
Subsense (ii)
(*Translation:* προσπάθεια, υποχρέωση.)
Synonyms: aim, attempt, document, drive‹1›, effort‹1›, energy‹1›, momentum‹1›, need, venture‹1›.
Related words: commitment, condition, debt, duty‹1›, obligation, responsibility‹1›, right‹1›.

capability
(*Translation:* ικανότητα, δυνατότητα.)
Synonyms: ability, capacity‹1›, competence, discretion‹1›, facility‹1›, power‹1›, qualification, resource‹1›, right‹1›, skill‹1›.
Related words: aptitude‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.

capacity
Subsense (i)
(*Translation:* ικανότητα, δυνατότητα, ιδιότητα, χαρακτηριστικό.)
Synonyms: ability, capability‹1›, characteristic, competence, discretion‹1›, facility‹1›, feature, parameter, particular, power‹1›, property‹2›, qualification, quality, resource‹1›, right‹1›, skill‹1›, status‹1›, trait.
Related words: aptitude‹1›, definition‹1›, designation, duty‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, venture‹1›.
Subsense (ii)
(*Translation:* δραστηριότητα.)
Synonyms: business‹1›, employment, exercise‹1›, experience, function‹1›, industry‹2›, instrument, intervention, labour, operating, study, work.

care
Sense 1
Subsense (i)
(*Translation:* προσοχή.)
Synonyms: importance.
Subsense (ii)
(*Translation:* θεραπεία.)
Synonyms: therapy.
Related words: rehabilitation‹1›, treating.
Sense 2
Subsense (i)
(*Translation:* φροντίδα.)
Synonyms: caring.
Subsense (ii)
(*Translation:* μέριμνα.)
Synonyms: responsibility‹1›.
Related words: auspice.

case
(*Translation:* ζήτημα, περίπτωση, διαδικασία.)
Hyperonyms: development‹1›.
Synonyms: argument, challenge‹1›, concern, dispute, factor‹1›, issue‹2›, judgment, matter, operation, option‹1›, point, problem, project, question, situation‹1›, sphere‹1›, subject‹1›, theme, topic.
Related words: aspect, chance, circumstance, condition, event‹1›, example, hypothesis, incident, likelihood, occasion, occurrence, opportunity, outbreak‹1›, possibility, probability, proceedings, scenario.

claim
(*Translation:* αξίωση, απαίτηση, αίτηση, αγωγή.)
Synonyms: appeal, applying‹1›, complaint, debt, demand, matter, need, obligation, petition, request‹1›, treatment‹1›.
Related words: constraint‹1›, entitlement‹1›, idea, necessity, plea‹1›, prerequisite, proceeding‹1›, proceedings, right‹1›, standard, taking‹2›, therapy, urgency.

competence
Hyperonyms: establishment‹1›.
Subsense (i)
(*Translation:* κατάρτιση, γνώση.)
Synonyms: background‹1›, communication, evidence, expertise‹2›, information, insight‹1›, know-how, knowledge, learning, prerequisite, proficiency, study, understanding‹1›.
Subsense (ii)
(*Translation:* αρμοδιότητα, δυνατότητα, ικανότητα.)
Synonyms: ability, aptitude‹1›, capability‹1›, capacity‹1›, discretion‹1›, facility‹1›, power‹1›, qualification, resource‹1›, right‹1›, skill‹1›.

Related words: assignment‹1›, ease, efficiency‹1›, expertise‹2›, function‹1›, jurisdiction‹1›, mandate, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, remit, responsibility‹1›, scope‹1›, status‹1›, venture‹1›.
Subsense (iii)
(*Translation:* αρχή.)
Synonyms: authority‹2›.
Related words: agency, association, beginning, body‹1›, commencement, concept‹1›, entity, policy‹1›, principle‹1›, right‹1›, rule‹1›, start, term, value.
conduct
Subsense (i)
(*Translation:* λειτουργία, υλοποίηση, διαχείριση, διεξαγωγή.)
Synonyms: achievement, achieving, application‹1›, attaining, attainment‹1›, birth, completion‹1›, creation‹1›, deployment‹2›, execution‹1›, functioning, governance, implementation, implementing, management‹1›, performing, preparation‹1›, progress, pursuance, pursuing, realisation, running, taking‹2›, task.
Related words: administering, effect‹1›, function‹1›, government, habit‹1›, handling, operating, performance‹1›, process, processing‹1›, purpose, service, significance, status‹1›, working.
Subsense (ii)
(*Translation:* πρακτική, άσκηση.)
Synonyms: accomplishment, commencing, demonstration‹1›, implementation, performance‹1›, practice, pursuit‹1›.
Related words: policy‹1›.
conquest
(*Translation:* άλωση.)
Synonyms: fall‹1›.
Related words: sack.
consideration
(*Translation:* εξέταση.)
Hyponyms: addressing, analysis, assessment‹1›, considering, consultation‹1›, debate, deliberation‹1›, discussing, discussion‹1›, examination‹1›, exploration, exploring, handling, hearing‹1›, inquiry, inspection, investigating‹1›, investigation‹1›, processing‹1›, reflection‹2›, screening‹1›, scrutiny‹1›, survey‹1›, tackling, test‹1›, thought, treatment‹1›.
consumption
Subsense (i)
(*Translation:* κατανάλωση.)
Synonyms: eating, intake‹3›.
Subsense (ii)
(*Translation:* χρήση.)
Synonyms: use.
Related words: applying‹1›, dosing‹2›, exploitation‹1›, management‹1›, operating, purpose, recourse‹1›, sourcing‹1›, taking‹2›, treatment‹1›, usage‹2›, using, work, working.
course
Hyperonyms: area.
Subsense (i)
(*Translation:* πλαίσιο.)
Synonyms: connection‹1›, context‹1›, realm, regime, scope‹1›.
Subsense (ii)
(*Translation:* πορεία.)
Synonyms: advance, evolution, process, progress, road, stage‹1›, state‹2›, walk‹1›, way.
Subsense (iii)
(*Translation:* ενέργεια.)
Synonyms: action.
Related words: act‹1›, attempt, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, force‹1›, goal‹1›, intervention, power‹1›, project, step‹1›.
creation
(*Translation:* υλοποίηση, ανάπτυξη, δημιουργία, διαμόρφωση, κατάρτιση.)
Hyperonyms: establishing‹1›, establishment‹1›.
Synonyms: achievement, achieving, attaining, attainment‹1›, birth, building‹1›, completion‹1›, conduct‹1›, developing‹1›, execution‹1›, formation, forming, formulating, implementation, implementing, introduction‹1›, manufacture, performing, preparation‹1›, production, pursuance, pursuing, realisation, setting-up, setting‹1›.
Related words: beginning, defining, definition‹1›, deployment‹1›, drafting‹1›, existence‹1›, formulation, framing, layout, preparing, producing, promotion‹1›.
culture
(*Translation:* κατάρτιση, μόρφωση.)
Hyperonyms: establishment‹1›.
Synonyms: learning.
Related words: knowledge, literacy, schooling, studies, teaching.
deliberation
(*Translation:* διαβούλευση, εξέταση, συζήτηση.)
Hyperonyms: consideration‹1›.
Synonyms: consultation‹1›, debate.
Related words: discussing.
delivery
(*Translation:* απόδοση, έκδοση, χορήγηση, παροχή.)
Synonyms: achievement, administering, allocation‹1›, awarding‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, paying‹1›, provision‹1›, publication‹1›, version‹1›.
Related words: administration‹1›, allocating, assigning‹1›, attribution, award‹1›, benefit‹1›, dosing‹2›, efficiency‹1›, issue‹2›, obligation, performance‹1›, providing‹1›, rendering, supply‹1›, therapy, treatment‹1›, use.
developing
(*Translation:* δημιουργία, ανάπτυξη, διαμόρφωση, κατάρτιση.)
Hyperonyms: establishing‹1›, establishment‹1›.
Synonyms: beginning, birth, building‹1›, completion‹1›, creation‹1›, deployment‹1›, existence‹1›, formation, forming, formulating, implementation, introduction‹1›, manufacture, production, promotion‹1›, setting-up, setting‹1›.
Related words: defining, definition‹1›, drafting‹1›, formulation, framing, layout, preparation‹1›, preparing, producing.
development
(*Translation:* διαδικασία.)
Hyponyms: act‹1›, action, activity‹1›, approach, arrangement‹1›, case‹1›, channel, configuration, court, energy‹1›, event‹1›, exercise‹1›, formality, hearing‹1›, incident, intervention, legislation, means, measure‹1›, mechanism, modality, network, operation, path, practice, procedure‹1›, proceeding‹1›, proceedings, process, progress, provision‹1›, route, rule‹1›, scheme, step‹1›, strategy, system‹1›, technique.
devising
Hyperonyms: establishing‹1›, establishment‹1›.
Subsense (i)

(*Translation:* κατάρτιση, ανάπτυξη, εκπόνηση, προπαρασκευή.)
Synonyms: drafting‹1›, preparing.
Related words: compilation, completion‹1›, producing.
Subsense (ii)
(*Translation:* σχεδιασμός.)
Synonyms: design‹1›, planning‹1›.
Subsense (iii)
(*Translation:* επεξεργασία.)
Synonyms: handling.
Related words: framing, processing‹1›, treating.

discretion
(*Translation:* ικανότητα, δυνατότητα.)
Synonyms: ability, capability‹1›, capacity‹1›, competence, facility‹1›, power‹1›, qualification, resource‹1›, right‹1›, skill‹1›.
Related words: aptitude‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.

discussion
(*Translation:* πρόβλημα, προβληματισμός, εξέταση, μελέτη, ανάλυση, εργασία.)
Hyperonyms: activity‹1›, consideration‹1›.
Synonyms: analysis, assessment‹1›, concern, considering, debate, discussing, problem, reflection‹2›, study, test‹1›, thought, trouble.
Related words: addressing, assay, challenge‹1›, condition, consultation‹1›, difficulty, dispute, document, exercise‹1›, gap, impact, interpretation, investigation‹1›, matter, obstacle‹1›, overview, project, reading‹1›, report, review, shortfall, survey‹1›, testing, trial.

dosing
Sense 1
(*Translation:* δόση.)
Synonyms: dose, dosage.
Sense 2
(*Translation:* χορήγηση, χρήση.)
Synonyms: provision‹1›, treatment‹1›, use.
Related words: administering, administration‹1›, allocation‹1›, applying‹1›, attribution, award‹1›, awarding‹1›, consumption‹2›, delivery‹1›, exploitation‹1›, giving‹1›, grant‹1›, granting‹1›, issue‹2›, issuing, management‹1›, operating, paying‹1›, purpose, recourse‹1›, sourcing‹1›, taking‹2›, therapy, usage‹2›, using, work, working.

drafting
(*Translation:* διατύπωση, κατάρτιση, διαμόρφωση, καθορισμός, εκπόνηση.)
Hyperonyms: establishment‹1›.
Synonyms: compilation, completion‹1›, configuration, defining, definition‹1›, design‹1›, devising, formality, formation, forming, formulating, formulation, framing, layout, preparing, producing, provision‹1›, setting‹1›, word, words.
Related words: birth, building‹1›, creation‹1›, determination‹2›, developing‹1›, establishing‹1›, guidance‹1›, introduction‹1›, manufacture, preparation‹1›, production, regulation, setting-up.

drive
Subsense (i)
(*Translation:* δυναμισμός.)
Synonyms: dynamism‹1›.
Subsense (ii)
(*Translation:* δράση, προσπάθεια, πρωτοβουλία.)
Synonyms: aim, approach, attempt, burden‹1›, effort‹1›, energy‹1›, incentive, initiative‹1›, momentum‹1›, need, operation, responsibility‹1›, venture‹1›.
Related words: action, effect‹1›, event‹1›, intervention, measure‹1›, potential‹1›, project, push, role‹1›, scheme, step‹1›, task, work.

duty
Sense 1
Subsense (i)
(*Translation:* επιταγή.)
Synonyms: command‹1›, guarantee, instruction‹2›.
Subsense (ii)
(*Translation:* υπηρεσία.)
Synonyms: service.
Related words: administration‹1›, agency, authority‹2›, body‹1›, centre, department, employment, facility‹1›, force‹1›, input‹2›, office‹1›, platform, point, sector, unit, work.
Subsense (iii)
(*Translation:* υποχρέωση, καθήκον.)
Synonyms: document, function‹1›, obligation, responsibility‹1›.
Related words: burden‹1›, commitment, condition, debt, mandate, mission, need, office‹1›, remit, right‹1›.
Subsense (iv)
(*Translation:* ιδιότητα.)
Synonyms: status‹1›.
Related words: capacity‹1›, characteristic, definition‹1›, designation, property‹2›, quality.
Sense 2
(*Translation:* τέλος.)
Synonyms: charge‹1›.
Related words: fee.

effect
Hyperonyms: arrangement‹1›.
Subsense (i)
(*Translation:* σκοπός.)
Synonyms: aim, end‹2›, focus‹1›, goal‹1›, intention, mission, objective, purpose, role‹1›, view.
Subsense (ii)
(*Translation:* λειτουργία, εφαρμογή.)
Synonyms: task.
Related words: application‹1›, conduct‹1›, function‹1›, functioning, governance, habit‹1›, implementing, management‹1›, operating, performance‹1›, process, purpose, running, service, significance, status‹1›, working.
Subsense (iii)
(*Translation:* ενέργεια, δράση.)
Synonyms: action, attempt, effort‹1›, energy‹1›, intervention, operation, project, step‹1›.
Related words: act‹1›, approach, course‹1›, drive‹1›, event‹1›, exercise‹1›, force‹1›, goal‹1›, incentive, initiative‹1›, measure‹1›, potential‹1›, power‹1›, push, responsibility‹1›, role‹1›, task, work.

efficiency
Subsense (i)
(*Translation:* επίδοση.)
Synonyms: attainment‹1›.
Subsense (ii)
(*Translation:* ικανότητα.)
Synonyms: qualification.
Related words: ability, aptitude‹1›, capability‹1›, capacity‹1›, competence, discretion‹1›,

performance‹1›, power‹1›, property‹2›, quality, resource‹1›, right‹1›, skill‹1›, status‹1›.
Subsense (iii)
(*Translation:* απόδοση.)
Synonyms: achievement.
Related words: allocating, assigning‹1›, delivery‹1›, rendering, version‹1›.
effort
(*Translation:* προσπάθεια, πρωτοβουλία, δράση, ενέργεια, εργασία.)
Hyperonyms: activity‹1›.
Synonyms: action, aim, approach, attempt, burden‹1›, drive‹1›, effect‹1›, energy‹1›, incentive, initiative‹1›, intervention, momentum‹1›, operation, project, responsibility‹1›, step‹1›, venture‹1›.
Related words: act‹1›, course‹1›, event‹1›, exercise‹1›, force‹1›, goal‹1›, measure‹1›, potential‹1›, power‹1›, push, role‹1›, scheme, task, work.
employment
Hyperonyms: activity‹1›.
Subsense (i)
(*Translation:* απόσπαση, δραστηριότητα, εργασία, απασχόληση.)
Synonyms: business‹1›, capacity‹1›, exercise‹1›, experience, function‹1›, industry‹2›, instrument, intervention, labour, occupation‹1›, operating, posting, profession‹1›, study, work.
Subsense (ii)
(*Translation:* υπηρεσία.)
Synonyms: service.
Related words: administration‹1›, agency, authority‹2›, body‹1›, centre, department, duty‹1›, facility‹1›, force‹1›, input‹2›, office‹1›, platform, point, sector, unit, work.
energy
(*Translation:* διαδικασία, δράση, προσπάθεια, ενέργεια, εργασία.)
Hyperonyms: activity‹1›, development‹1›.
Synonyms: action, aim, attempt, burden‹1›, drive‹1›, effect‹1›, effort‹1›, intervention, momentum‹1›, need, operation, project, step‹1›, venture‹1›.
Related words: act‹1›, approach, course‹1›, event‹1›, exercise‹1›, force‹1›, goal‹1›, incentive, initiative‹1›, measure‹1›, potential‹1›, power‹1›, push, responsibility‹1›, role‹1›, task, work.
enhancing
(*Translation:* ενδυνάμωση, αύξηση, βελτίωση, προώθηση, ενίσχυση.)
Synonyms: assistance‹1›, enhancement, improvement, improving‹1›, increase‹1›, increasing, promotion‹1›, raising‹1›, strengthening, support.
Related words: adopting, aid‹1›, assisting, encouraging, expansion, facilitating, fostering, generating, grant‹1›, growth‹1›, help, launch, launching, proliferation, promoting‹1›, rise, supporting.
establishment
(*Translation:* κατάρτιση.)
Hyponyms: competence, creation‹1›, culture‹1›, defining, definition‹1›, developing‹1›, devising, drafting‹1›, education, establishing‹1›, expertise‹2›, formation, forming, formulating, framing, layout, learning, preparation‹1›, preparing, producing, qualification, schooling, setting‹1›, studies, teaching, training‹1›.
event
Hyperonyms: development‹1›.
Subsense (i)
(*Translation:* έλεγχος.)
Hyponyms: assessment‹1›, audit, check, control‹1›, detecting, examination‹1›, inquiry, inspection, investigating‹1›, report, research, review, screening‹1›, scrutiny‹1›, search, searching, service, servicing, supervision, survey‹1›, test‹1›, testing, verification.
Subsense (ii)
(*Translation:* περίπτωση, διαδικασία, δράση.)
Synonyms: operation.
Related words: action, approach, aspect, attempt, case‹1›, chance, circumstance, condition, drive‹1›, effect‹1›, effort‹1›, energy‹1›, example, hypothesis, incentive, incident, initiative‹1›, intervention, likelihood, matter, measure‹1›, occasion, occurrence, opportunity, outbreak‹1›, possibility, potential‹1›, probability, proceedings, project, push, responsibility‹1›, role‹1›, scenario, situation‹1›, step‹1›, task, work.
examination
(*Translation:* έρευνα, έλεγχος, εξέταση.)
Hyperonyms: event‹1›, consideration‹1›.
Hyponyms: analysis, audit, check, control‹1›, detecting, exploration, exploring, inquiry, investigating‹1›, investigation‹1›, research, review, screening‹1›, scrutiny‹1›, search, searching, study, survey‹1›, trial, use, verification, work.
excursion
Subsense (i)
(*Translation:* εκδρομή.)
Synonyms: trip.
Subsense (ii)
(*Translation:* μετακίνηση.)
Synonyms: shift‹1›.
Related words: journey‹1›, mobility‹1›, move‹1›, travelling.
exercise
Hyperonyms: activity‹1›, development‹1›.
Subsense (i)
(*Translation:* διαδικασία, μελέτη, ενέργεια, εργασία, δραστηριότητα.)
Synonyms: action, assessment‹1›, business‹1›, capacity‹1›, employment, experience, function‹1›, industry‹2›, instrument, intervention, labour, operating, project, study.
Related words: act‹1›, analysis, assay, attempt, consultation‹1›, course‹1›, debate, discussing, discussion‹1›, document, effect‹1›, effort‹1›, energy‹1›, force‹1›, goal‹1›, investigation‹1›, power‹1›, reading‹1›, reflection‹2›, report, step‹1›, survey‹1›, trial.
Subsense (ii)
(*Translation:* εκτέλεση.)
Synonyms: preparation‹1›.
Related words: accomplishment, achieving, attaining, completion‹1›, enforcement, execution‹1›, fulfilling‹1›, implementation, implementing, performance‹1›, performing, pursuance.
expertise
Hyperonyms: establishment‹1›.
Subsense (i)
(*Translation:* εμπειρία.)
Synonyms: experience, scheme.
Subsense (ii)
(*Translation:* γνώση, κατάρτιση.)

Synonyms: background‹1›, communication, competence, evidence, information, insight‹1›, know-how, knowledge, learning, prerequisite, proficiency, skill‹1›, study, understanding‹1›.
Subsense (iii)
(*Translation:* αρμοδιότητα.)
Synonyms: power‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, function‹1›, jurisdiction‹1›, mandate, remit, responsibility‹1›, scope‹1›.

exploitation
(*Translation:* χρήση, αξιοποίηση, εκμετάλλευση.)
Synonyms: exploiting, harnessing, operating, usage‹2›, use, using, utilisation.
Related words: applying‹1›, command‹1›, consumption‹2›, dosing‹2›, management‹1›, purpose, recourse‹1›, sourcing‹1›, taking‹2›, treatment‹1›, upgrading, work, working.

facility
Subsense (i)
(*Translation:* δυνατότητα.)
Synonyms: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, ease, option‹1›, possibility, potential‹1›, power‹1›, resource‹1›, right‹1›, skill‹1›, venture‹1›.
Subsense (ii)
(*Translation:* χώρος, υπηρεσία.)
Synonyms: area, building‹1›, chamber, context‹1›, environment, field‹1›, location‹1›, office‹1›, part, place, region, sector, service, setting‹1›, site, surroundings, terrain.
Related words: administration‹1›, agency, authority‹2›, body‹1›, centre, department, duty‹1›, employment, force‹1›, input‹2›, platform, point, unit, work.

factor
Hyperonyms: part.
Subsense (i)
(*Translation:* συνθήκη.)
Synonyms: agreement, background‹2›, circumstance, condition, context‹1›, convention, manner, occasion, situation‹1›, state‹1›, treaty.
Subsense (ii)
(*Translation:* ζήτημα, θέμα, στοιχείο.)
Synonyms: case‹1›, issue‹2›, matter, option‹1›, point, problem, question, sphere‹1›, subject‹1›, theme, topic.
Related words: affair‹1›, aim, area, argument, aspect, challenge‹1›, concern, discipline‹1›, dispute, fact, field‹1›, incident, item, judgment, project, situation‹1›, unit.

fall
(*Translation:* άλωση.)
Synonyms: conquest‹1›, sack.

flow
Subsense (i)
(*Translation:* ροή.)
Synonyms: stream‹2›.
Subsense (ii)
(*Translation:* διάδοση, κυκλοφορία, διακίνηση.)
Synonyms: circulation, disseminating, marketing‹1›, mobility‹1›, movement‹1›, sharing‹1›, trade‹1›, traffic‹1›, transfer‹1›, transit, transport‹1›, transportation, travel‹1›.
Related words: demonstration‹1›, dissemination, distribution, expansion, proliferation, promotion‹1›, propagation, spread, transmission.

focus
Subsense (i)
(*Translation:* σκοπός, στόχος.)
Synonyms: aim, aspiration, effect‹1›, end‹2›, goal‹1›, idea, intention, mission, need, objective, purpose, scope‹1›, vision‹1›.
Related words: role‹1›, view.
Subsense (ii)
(*Translation:* αναφορά.)
Synonyms: provision‹1›.
Related words: clause, communication, complaint, contribution, declaration, document, identification‹1›, indication, mention, overview, petition, reference‹1›, regard‹1›, report, review, statement.

force
Subsense (i)
(*Translation:* υπηρεσία, φορέας.)
Synonyms: actor‹1›, agency, agent, body‹1›, entity, facilitator, institute, institution, operator, organisation‹1›, partner, party, player, service, structure‹1›.
Related words: administration‹1›, authority‹2›, centre, department, duty‹1›, employment, facility‹1›, input‹2›, office‹1›, platform, point, sector, unit, work.
Subsense (ii)
(*Translation:* ενέργεια.)
Synonyms: action.
Related words: act‹1›, attempt, course‹1›, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, goal‹1›, intervention, power‹1›, project, step‹1›.

framing
Hyperonyms: establishment‹1›.
Subsense (i)
(*Translation:* καθορισμός, χάραξη, διαμόρφωση, κατάρτιση.)
Synonyms: defining, definition‹1›, drafting‹1›, formation, forming, formulating, formulation, layout, producing, provision‹1›, setting‹1›.
Related words: adoption, birth, building‹1›, creation‹1›, design‹1›, determination‹2›, developing‹1›, establishing‹1›, guidance‹1›, introduction‹1›, manufacture, preparation‹1›, preparing, production, regulation, setting-up.
Subsense (ii)
(*Translation:* επεξεργασία.)
Synonyms: handling.
Related words: devising, processing‹1›, treating.

function
Subsense (i)
(*Translation:* αρμοδιότητα, καθήκον.)
Synonyms: duty‹1›, mandate, mission, obligation, office‹1›, power‹1›, remit, responsibility‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, expertise‹2›, jurisdiction‹1›, scope‹1›.
Subsense (ii)
(*Translation:* λειτουργία.)
Synonyms: task.
Related words: application‹1›, conduct‹1›, effect‹1›, functioning, governance, habit‹1›, implementing, management‹1›, operating, performance‹1›, process, purpose, running, service, significance, status‹1›, working.
Subsense (iii)
(*Translation:* δραστηριότητα.)
Synonyms: business‹1›, capacity‹1›, employment, exercise‹1›, experience, industry‹2›, instrument, intervention, labour, operating, study, work.

giving
Subsense (i)
(*Translation:* ανάθεση, παροχή, προσφορά, χορήγηση.)
Synonyms: administering, allocation‹1›, attribution, award‹1›, awarding‹1›, contribution, delivery‹1›, donation, grant‹1›, granting‹1›, issuing, opening‹2›, paying‹1›, provision‹1›, supply‹1›.
Related words: administration‹1›, attachment‹1›, benefit‹1›, conferring, contracting, delegation‹1›, dosing‹2›, issue‹2›, obligation, performance‹1›, providing‹1›, therapy, treatment‹1›, use.
Subsense (ii)
(*Translation:* θέσπιση.)
Synonyms: introduction‹1›.
Related words: adopting, adoption, forming, introducing, launch, launching, taking‹1›.
goal
Subsense (i)
(*Translation:* ζητούμενο.)
Synonyms: question.
Subsense (ii)
(*Translation:* ενέργεια.)
Synonyms: action.
Related words: act‹1›, attempt, course‹1›, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, force‹1›, intervention, power‹1›, project, step‹1›.
Subsense (iii)
(*Translation:* σκοπός, στόχος.)
Synonyms: aim, effect‹1›, end‹2›, focus‹1›, intention, mission, objective, purpose.
Related words: aspiration, idea, need, role‹1›, scope‹1›, view, vision‹1›.
grant
Subsense (i)
(*Translation:* επιδότηση.)
Synonyms: fund, funding.
Subsense (ii)
(*Translation:* λήψη, παροχή, χορήγηση.)
Synonyms: administering, administration‹1›, allocation‹1›, awarding‹1›, delivery‹1›, giving‹1›, granting‹1›, issuing, paying‹1›, provision‹1›.
Related words: attribution, award‹1›, benefit‹1›, dosing‹2›, intake‹3›, issue‹2›, obligation, obtaining‹1›, performance‹1›, providing‹1›, receipt, receiving, reception‹1›, sourcing‹1›, supply‹1›, taking‹2›, therapy, treatment‹1›, use.
Subsense (iii)
(*Translation:* ενίσχυση.)
Synonyms: assistance‹1›.
Related words: aid‹1›, enhancement, enhancing‹1›, generating, improvement, improving‹1›, increase‹1›, increasing, strengthening, support.
granting
(*Translation:* χορήγηση, έκδοση, παροχή.)
Synonyms: administering, allocation‹1›, awarding‹1›, delivery‹1›, giving‹1›, grant‹1›, issuing, paying‹1›, provision‹1›.
Related words: administration‹1›, attribution, award‹1›, benefit‹1›, dosing‹2›, issue‹2›, obligation, performance‹1›, providing‹1›, publication‹1›, supply‹1›, therapy, treatment‹1›, use, version‹1›.
hearing
(*Translation:* δίκη.)
Hyperonyms: consideration‹1›, development‹1›.
Synonyms: court, litigation.

improving
Subsense (i)
(*Translation:* προαγωγή.)
Synonyms: encouraging, facilitating, fostering, promoting‹1›.
Subsense (ii)
(*Translation:* ενίσχυση, βελτίωση, αύξηση.)
Synonyms: assistance‹1›, enhancement, enhancing‹1›, improvement, increase‹1›, increasing, raising‹1›.
Related words: aid‹1›, expansion, generating, grant‹1›, growth‹1›, proliferation, promotion‹1›, rise, strengthening, support.
increase
(*Translation:* βελτίωση, ενίσχυση, αύξηση, ανάπτυξη.)
Hyperonyms: establishing‹1›.
Synonyms: increasing, assistance‹1›, enhancement, enhancing‹1›, improvement, improving‹1›, promotion‹1›, raising‹1›.
Related words: aid‹1›, expansion, generating, grant‹1›, growth‹1›, proliferation, rise, strengthening, support.
industry
Subsense (i)
(*Translation:* τομέας, κλάδος.)
Synonyms: branch, context‹1›, discipline‹1›, profession‹1›, trade‹2›.
Related words: connection‹1›, department, environment, field‹1›, realm, section, sector, sphere‹1›, subject‹1›, theme, topic.
Subsense (ii)
(*Translation:* επιχείρηση, μονάδα, δραστηριότητα.)
Synonyms: agency, block‹1›, business‹1›, capacity‹1›, employment, exercise‹1›, experience, function‹1›, instrument, intervention, labour, operating, piece, plant‹1›, point, study, unit, work.
Related words: company‹1›, corporation, enterprise, firm, start-up, undertaking‹1›, venture‹1›.
initiation
(*Translation:* κίνηση, έναρξη, εισαγωγή.)
Synonyms: beginning, commencement, commencing, introducing, launch, launching, opening‹1›, start, starting.
Related words: admission‹1›, entrance, entry, input‹1›, intake‹1›.
input
Sense 1
(*Translation:* είσοδος, εισαγωγή.)
Synonyms: intake‹1›, admission‹1›, entrance, entry, introducing.
Related words: initiation‹1›, starting.
Sense 2
Subsense (i)
(*Translation:* υπηρεσία.)
Synonyms: service.
Related words: administration‹1›, agency, authority‹2›, body‹1›, centre, department, duty‹1›, employment, facility‹1›, force‹1›, office‹1›, platform, point, sector, unit, work.
Subsense (ii)
(*Translation:* ισχύς.)
Synonyms: strength‹1›.
Related words: validity‹1›.
Sense 3
(*Translation:* συνδρομή.)
Synonyms: aid‹1›.

Related words: assisting, contribution, donation, help, helping, support.
inspection
Hyperonyms: event‹1›, consideration‹1›.
Subsense (i)
(*Translation:* επαλήθευση.)
Synonyms: checking, verification.
Subsense (ii)
(*Translation:* επιθεώρηση, έλεγχος, εξέταση.)
Synonyms: supervision.
installation
Subsense (i)
(*Translation:* συσκευή.)
Synonyms: device‹1›.
Related words: apparatus, appliance, equipment.
Subsense (ii)
(*Translation:* εγκατάσταση.)
Synonyms: plant‹1›.
Related words: smelter, station‹1›.
introduction
Hyperonyms: arrangement‹1›.
Subsense (i)
(*Translation:* εφαρμογή, θέσπιση.)
Synonyms: adopting, adoption, forming, giving‹1›, introducing, launch, launching, taking‹1›.
Subsense (ii)
(*Translation:* διαμόρφωση, δημιουργία.)
Synonyms: birth, building‹1›, creation‹1›, developing‹1›, formation, forming, formulating, manufacture, production, setting-up, setting‹1›.
Related words: beginning, completion‹1›, defining, definition‹1›, deployment‹1›, drafting‹1›, existence‹1›, formulation, framing, implementation, layout, preparation‹1›, preparing, producing, promotion‹1›.
investigating
(*Translation:* εξέταση, έρευνα, διερεύνηση, έλεγχος.)
Hyperonyms: examination‹1›, event‹1›, consideration‹1›.
Synonyms: survey‹1›.
Related words: exploration, exploring, inquiry, investigation‹1›, search.
investigation
(*Translation:* εξέταση, μελέτη, έρευνα, ανίχνευση, διερεύνηση.)
Hyperonyms: examination‹1›, consideration‹1›.
Synonyms: assessment‹1›, survey‹1›, testing.
Related words: analysis, assay, consultation‹1›, debate, detecting, detection, discussing, discussion‹1›, document, exercise‹1›, exploration, exploring, inquiry, investigating‹1›, project, reading‹1›, recognition, reflection‹2›, report, screening‹1›, search, study, trial.
issue
Subsense (i)
(*Translation:* ζήτημα, θέμα.)
Synonyms: affair‹1›, aim, area, aspect, case‹1›, discipline‹1›, fact, factor‹1›, field‹1›, incident, item, matter, option‹1›, point, problem, question, sphere‹1›, subject‹1›, theme, topic, unit.
Related words: argument, challenge‹1›, concern, dispute, judgment, project, situation‹1›.
Subsense (ii)
(*Translation:* χορήγηση.)
Synonyms: provision‹1›.
Related words: administering, administration‹1›, allocation‹1›, attribution, award‹1›, awarding‹1›, delivery‹1›, dosing‹2›, giving‹1›, grant‹1›, granting‹1›, issuing, paying‹1›, therapy, treatment‹1›, use.
job
Hyperonyms: activity‹1›.
Subsense (i)
(*Translation:* αποστολή.)
Synonyms: consignment, mission, posting, undertaking‹1›.
Subsense (ii)
(*Translation:* θέση.)
Synonyms: role‹1›.
Related words: allocation‹1›, attention, location‹1›, office‹1›, opinion, place, placing, position, post‹1›, release‹1›, seat, settlement‹1›, site, situation‹1›, spot, status‹1›.
Subsense (iii)
(*Translation:* δουλειά.)
Synonyms: practice.
journey
(*Translation:* μετακίνηση, ταξίδι.)
Synonyms: shift‹1›, tour, travelling.
Related words: excursion‹1›, mobility‹1›, move‹1›, trip.
jurisdiction
(*Translation:* αρμοδιότητα.)
Synonyms: power‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, expertise‹2›, function‹1›, mandate, remit, responsibility‹1›, scope‹1›.
management
(*Translation:* διοίκηση, χρήση, λειτουργία, διαχείριση, αντιμετώπιση.)
Synonyms: administering, command‹1›, conduct‹1›, governance, government, handling, operating, processing‹1›, purpose, running, task, treatment‹1›, use, working.
Related words: addressing, application‹1›, applying‹1›, consumption‹2›, countering, dosing‹2›, effect‹1›, exploitation‹1›, function‹1›, functioning, habit‹1›, implementing, meeting‹2›, performance‹1›, preventing, process, recourse‹1›, response‹3›, service, settlement‹1›, significance, sourcing‹1›, status‹1›, tackling, taking‹2›, therapy, treating, usage‹2›, using, work.
mandate
Subsense (i)
(*Translation:* αίτημα.)
Synonyms: appeal, demand, motion, request‹1›.
Subsense (ii)
(*Translation:* αρμοδιότητα, καθήκον.)
Synonyms: function‹1›, power‹1›, remit, responsibility‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, duty‹1›, expertise‹2›, jurisdiction‹1›, mission, obligation, office‹1›, scope‹1›.
marketing
Subsense (i)
(*Translation:* εμπορία.)
Synonyms: market, trading, trafficking.
Subsense (ii)
(*Translation:* διάθεση.)
Synonyms: dissemination.
Related words: allocating, availability‹1›, disclosure, distribution, placing, possession‹1›, release‹1›, supply‹1›, taking‹1›.
Subsense (iii)
(*Translation:* κυκλοφορία.)

Synonyms: circulation, flow‹1›, mobility‹1›, movement‹1›, sharing‹1›, traffic‹1›, transit, travel‹1›.
meeting
Sense 1
(*Translation:* σύνοδος.)
Synonyms: assembly, session.
Sense 2
Subsense (i)
(*Translation:* ικανοποίηση.)
Synonyms: enforcement, fulfillment.
Subsense (ii)
(*Translation:* αντιμετώπιση.)
Synonyms: treatment‹1›.
Related words: addressing, countering, handling, management‹1›, preventing, processing‹1›, response‹3›, settlement‹1›, tackling, therapy, treating.
Subsense (iii)
(*Translation:* εκπλήρωση, επίτευξη.)
Synonyms: achievement, completion‹1›, fulfilling‹1›.
Related words: accomplishment, achieving, attaining, attainment‹1›, compliance‹1›, execution‹1›, fulfilment, pursuing, pursuit‹1›, realisation.
Sense 3
(*Translation:* εκδήλωση.)
Synonyms: initiative‹1›.
Related words: contracting, expression‹1›, feast, manifestation.
mission
Subsense (i)
(*Translation:* στόχος, σκοπός.)
Synonyms: aim, effect‹1›, end‹2›, focus‹1›, goal‹1›, intention, objective, purpose.
Related words: aspiration, idea, need, role‹1›, scope‹1›, view, vision‹1›.
Subsense (ii)
(*Translation:* καθήκον.)
Synonyms: function‹1›.
Related words: duty‹1›, mandate, obligation, office‹1›, remit, responsibility‹1›.
Subsense (iii)
(*Translation:* αποστολή.)
Synonyms: job.
Related words: consignment, posting, undertaking‹1›.
mobility
Subsense (i)
(*Translation:* κυκλοφορία, διακίνηση.)
Synonyms: circulation, disseminating, flow‹1›, marketing‹1›, movement‹1›, sharing‹1›, trade‹1›, traffic‹1›, transit, transport‹1›, transportation, travel‹1›.
Subsense (ii)
(*Translation:* μετακίνηση.)
Synonyms: shift‹1›.
Related words: excursion‹1›, journey‹1›, move‹1›, travelling.
momentum
Subsense (i)
(*Translation:* ώθηση.)
Synonyms: push.
Subsense (ii)
(*Translation:* προσπάθεια.)
Synonyms: aim, attempt, burden‹1›, drive‹1›, effort‹1›, energy‹1›, need, venture‹1›.
move
Subsense (i)
(*Translation:* μετάβαση.)
Synonyms: transition.
Subsense (ii)
(*Translation:* κίνηση.)
Synonyms: traffic‹1›.
Related words: organisation‹2›, trend‹1›.
Subsense (iii)
(*Translation:* μετακίνηση.)
Synonyms: shift‹1›.
Related words: excursion‹1›, journey‹1›, mobility‹1›, travelling.
movement
(*Translation:* διακίνηση, κυκλοφορία.)
Synonyms: circulation, flow‹1›, marketing‹1›, mobility‹1›, sharing‹1›, trade‹1›, traffic‹1›, transportation.
obligation
Subsense (i)
(*Translation:* απαίτηση, καθήκον, υποχρέωση, κανόνας.)
Synonyms: claim‹1›, constraint‹1›, convention, debt, demand, directive, document, duty‹1›, entitlement‹1›, function‹1›, legislation, necessity, need, policy‹1›, prerequisite, principle‹1›, regulation, request‹1›, responsibility‹1›, rule‹1›, standard, urgency.
Related words: burden‹1›, commitment, condition, mandate, mission, office‹1›, remit, right‹1›.
Subsense (ii)
(*Translation:* λύση.)
Synonyms: option‹1›.
Related words: answer, approach, decision, determination‹1›, device‹1›, judgment, means, possibility, reasoning‹1›, remedy‹1›, resort, response‹3›, result‹1›, settlement‹1›, solution, solving, step‹1›.
Subsense (iii)
(*Translation:* παροχή.)
Synonyms: allocation‹1›.
Related words: administering, awarding‹1›, benefit‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, paying‹1›, performance‹1›, providing‹1›, supply‹1›.
occupation
Hyperonyms: activity‹1›.
Subsense (i)
(*Translation:* επάγγελμα.)
Synonyms: trade‹2›.
Subsense (ii)
(*Translation:* εργασία, απασχόληση.)
Synonyms: employment.
Related words: labour, profession‹1›.
office
Subsense (i)
(*Translation:* υπηρεσία, θέση, κέντρο, χώρος, οργανισμός.)
Synonyms: administration‹1›, agency, body‹1›, centre, entity, facility‹1›, institution, location‹1›, organisation‹1›, place, role‹1›, sector, service, settlement‹1›, site.
Related words: allocation‹1›, area, attention, authority‹2›, building‹1›, chamber, context‹1›, department, duty‹1›, employment, environment, field‹1›, force‹1›, foundation‹1›, input‹2›, job, middle, opinion, part, placing, platform, point, position, post‹1›, region, release‹1›, resort, seat, setting‹1›, shop, situation‹1›, spot, status‹1›, surroundings, terrain, unit, work.
Subsense (ii)

(*Translation:* καθήκον.)
Synonyms: function‹1›.
Related words: duty‹1›, mandate, mission, obligation, remit, responsibility‹1›.

opening
Sense 1
(*Translation:* έναρξη, κίνηση.)
Synonyms: beginning, commencement, commencing, initiation‹1›, launch, launching, start, starting.
Sense 2
Subsense (i)
(*Translation:* προσφορά.)
Synonyms: contribution, donation, giving‹1›, supply‹1›.
Subsense (ii)
(*Translation:* ίδρυση.)
Synonyms: building‹1›.
Related words: start-up.

option
Subsense (i)
(*Translation:* λύση.)
Synonyms: answer, approach, decision, determination‹1›, device‹1›, judgment, means, obligation, possibility, reasoning‹1›, remedy‹1›, resort, response‹3›, result‹1›, settlement‹1›, solution, solving, step‹1›.
Subsense (ii)
(*Translation:* άποψη, θέμα, ζήτημα.)
Synonyms: argument, aspect, case‹1›, concern, factor‹1›, issue‹2›, matter, point, problem, question, sphere‹1›, subject‹1›, theme, topic.
Related words: affair‹1›, aim, angle, approach, area, assumption‹1›, challenge‹1›, contribution, demand, direction‹1›, discipline‹1›, dispute, fact, field‹1›, idea, incident, item, judgment, message, opinion, perspective, position, presumption, project, side‹1›, situation‹1›, statement, unit, view, viewpoint.
Subsense (iii)
(*Translation:* δυνατότητα.)
Synonyms: facility‹1›.
Related words: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, ease, possibility, potential‹1›, power‹1›, resource‹1›, right‹1›, skill‹1›, venture‹1›.

organisation
Sense 1
(*Translation:* διοργάνωση, οργανισμός, φορέας, οργάνωση.)
Synonyms: agency, body‹1›, entity, force‹1›, institution, office‹1›, organising, structure‹1›.
Related words: actor‹1›, administration‹1›, agent, facilitator, institute, launch, operator, partner, party, planning‹1›, player.
Sense 2
(*Translation:* κίνηση.)
Synonyms: traffic‹1›.
Related words: move‹1›, trend‹1›.

paying
(*Translation:* χορήγηση, παροχή.)
Synonyms: administering, allocation‹1›, awarding‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, provision‹1›.
Related words: administration‹1›, attribution, award‹1›, benefit‹1›, dosing‹2›, issue‹2›, obligation, performance‹1›, providing‹1›, supply‹1›, therapy, treatment‹1›, use.

performance
Subsense (i)
(*Translation:* εκτέλεση, άσκηση.)
Synonyms: accomplishment, commencing, conduct‹1›, demonstration‹1›, practice, preparation‹1›, pursuit‹1›.
Related words: achieving, attaining, completion‹1›, enforcement, execution‹1›, exercise‹1›, fulfilling‹1›, implementation, implementing, performing, pursuance.
Subsense (ii)
(*Translation:* αξιολόγηση.)
Synonyms: analysis, appraisal‹1›, assessing, considering, control‹1›, evaluating, evaluation, interpreting, monitoring‹1›, review, testing, valuation, valuing‹1›.
Subsense (iii)
(*Translation:* λειτουργία.)
Synonyms: task.
Related words: application‹1›, conduct‹1›, effect‹1›, function‹1›, functioning, governance, habit‹1›, implementing, management‹1›, operating, process, purpose, running, service, significance, status‹1›, working.
Subsense (iv)
(*Translation:* ικανότητα.)
Synonyms: qualification.
Related words: ability, aptitude‹1›, capability‹1›, capacity‹1›, competence, discretion‹1›, efficiency‹1›, power‹1›, property‹2›, quality, resource‹1›, right‹1›, skill‹1›, status‹1›.
Subsense (v)
(*Translation:* παροχή.)
Synonyms: allocation‹1›.
Related words: administering, awarding‹1›, benefit‹1›, delivery‹1›, giving‹1›, grant‹1›, granting‹1›, issuing, obligation, paying‹1›, providing‹1›, supply‹1›.

petition
Subsense (i)
(*Translation:* αναφορά.)
Synonyms: provision‹1›.
Related words: clause, communication, complaint, contribution, declaration, document, focus‹1›, identification‹1›, indication, mention, overview, reference‹1›, regard‹1›, report, review, statement.
Subsense (ii)
(*Translation:* αγωγή, προσφυγή, αίτηση, διαφορά.)
Synonyms: appeal, claim‹1›, complaint, proceeding‹1›, proceedings, review, treatment‹1›.
Related words: applying‹1›, demand, difficulty, dispute, gap, litigation, matter, plea‹1›, problem, recourse‹1›, remedy‹1›, request‹1›, taking‹2›, therapy.

plant
Subsense (i)
(*Translation:* εγκατάσταση.)
Synonyms: installation, smelter, station‹1›.
Subsense (ii)
(*Translation:* μονάδα.)
Synonyms: agency, block‹1›, industry‹2›, piece, point, unit.

plea
Subsense (i)
(*Translation:* αγωγή.)
Synonyms: treatment‹1›.
Related words: appeal, claim‹1›, petition, proceeding‹1›, proceedings, taking‹2›, therapy.
Subsense (ii)

(*Translation:* λόγος, ισχυρισμός.)
Synonyms: argument, purpose.
Related words: fact, ground‹1›, need, reason‹1›, word.
possession
(*Translation:* ύπαρξη, διάθεση, αγαθό, ιδιοκτησία.)
Synonyms: availability‹1›, dissemination, existence‹1›, presence, property‹2›, right‹1›, supply‹1›.
Related words: allocating, consignment, disclosure, distribution, marketing‹1›, ownership, placing, possessions, principle‹1›, release‹1›, taking‹1›.
possibility
Subsense (i)
(*Translation:* πρόταση, λύση, μέσο.)
Synonyms: aid‹1›, decision, directive, draft, means, middle, opinion, option‹1›, proposal, recommendation, regulation, remedy‹1›, resort, sentence, solution, statement, submission, suggestion, tool‹1›, way.
Related words: answer, approach, determination‹1›, device‹1›, judgment, obligation, reasoning‹1›, response‹3›, result‹1›, settlement‹1›, solving, step‹1›.
Subsense (ii)
(*Translation:* περίπτωση.)
Synonyms: operation.
Related words: aspect, case‹1›, chance, circumstance, condition, event‹1›, example, hypothesis, incident, likelihood, matter, occasion, occurrence, opportunity, outbreak‹1›, probability, proceedings, scenario, situation‹1›.
Subsense (iii)
(*Translation:* δυνατότητα.)
Synonyms: facility‹1›.
Related words: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, ease, option‹1›, potential‹1›, power‹1›, resource‹1›, right‹1›, skill‹1›, venture‹1›.
post
(*Translation:* θέση.)
Synonyms: role‹1›.
Related words: allocation‹1›, attention, job, location‹1›, office‹1›, opinion, place, placing, position, release‹1›, seat, settlement‹1›, site, situation‹1›, spot, status‹1›.
potential
Subsense (i)
(*Translation:* πιθανότητα.)
Synonyms: chance, likelihood, probability, prospect, risk.
Subsense (ii)
(*Translation:* δράση.)
Synonyms: operation.
Related words: action, approach, attempt, drive‹1›, effect‹1›, effort‹1›, energy‹1›, event‹1›, incentive, initiative‹1›, intervention, measure‹1›, project, push, responsibility‹1›, role‹1›, step‹1›, task, work.
Subsense (iii)
(*Translation:* δυνατότητα.)
Synonyms: facility‹1›.
Related words: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, ease, option‹1›, possibility, power‹1›, resource‹1›, right‹1›, skill‹1›, venture‹1›.
power
Subsense (i)
(*Translation:* δυνατότητα, ικανότητα, αρμοδιότητα.)
Synonyms: ability, aptitude‹1›, assignment‹1›, capability‹1›, capacity‹1›, competence, discretion‹1›, expertise‹2›, facility‹1›, function‹1›, jurisdiction‹1›, mandate, qualification, remit, resource‹1›, responsibility‹1›, right‹1›, scope‹1›, skill‹1›.
Related words: ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.
Subsense (ii)
(*Translation:* ενέργεια.)
Synonyms: action.
Related words: act‹1›, attempt, course‹1›, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, force‹1›, goal‹1›, intervention, project, step‹1›.
preparation
(*Translation:* εκτέλεση, ανάπτυξη, υλοποίηση, διαμόρφωση, κατάρτιση.)
Hyperonyms: establishing‹1›, establishment‹1›.
Synonyms: accomplishment, achievement, achieving, application‹1›, attaining, attainment‹1›, birth, completion‹1›, conduct‹1›, creation‹1›, enforcement, execution‹1›, exercise‹1›, fulfilling‹1›, implementation, implementing, performance‹1›, performing, pursuance, pursuing, realisation, setting‹1›.
Related words: building‹1›, defining, definition‹1›, developing‹1›, drafting‹1›, formation, forming, formulating, formulation, framing, introduction‹1›, layout, manufacture, preparing, producing, production, setting-up.
prescribing
(*Translation:* χορήγηση.)
Synonyms: prescription‹1›.
prescription
(*Translation:* χορήγηση.)
Synonyms: prescribing‹1›.
procedure
(*Translation:* μέθοδος, εργασία, διαδικασία, διάταξη.)
Hyperonyms: activity‹1›, development‹1›.
Synonyms: approach, arrangements, form‹1›, means, method, modality, mode‹1›, pattern, practice, process, requirement, route, solution, strategy, system‹1›, technique, test‹1›, way.
Related words: clause, configuration, device‹1›, ground‹1›, instrument, legislation, measure‹1›, order‹1›, principle‹1›, regulation, rule‹1›, scheme, step‹1›.
proceeding
(*Translation:* διαφορά, προσφυγή, αγωγή, δίκη.)
Hyperonyms: development‹1›.
Synonyms: appeal, court, litigation, petition, proceedings, review, treatment‹1›.
Related words: claim‹1›, complaint, difficulty, dispute, gap, matter, plea‹1›, problem, recourse‹1›, remedy‹1›, taking‹2›, therapy.
processing
Sense 1
(*Translation:* επεξεργασία, αντιμετώπιση, διαχείριση, εξέταση.)
Hyperonyms: consideration‹1›.
Synonyms: handling, devising, framing, management‹1›, treating, treatment‹1›.
Related words: addressing, administering, conduct‹1›, countering, governance, government, meeting‹2›, preventing, response‹3›, settlement‹1›, tackling, therapy.
Sense 2
(*Translation:* μεταποίηση, βιομηχανία.)
Synonyms: manufacture, manufacturing.

profession
Subsense (i)
(*Translation:* κλάδος, επάγγελμα.)
Synonyms: industry‹2›, trade‹2›.
Related words: branch, discipline‹1›.
Subsense (ii)
(*Translation:* απασχόληση.)
Synonyms: employment.
Related words: labour, occupation‹1›.
project
Hyperonyms: activity‹1›.
Subsense (i)
(*Translation:* εργασία, ενέργεια, δράση, μελέτη, έργο.)
Synonyms: achievement, action, analysis, assessment‹1›, attempt, contribution, effect‹1›, effort‹1›, energy‹1›, exercise‹1›, intervention, operation, record‹1›, step‹1›.
Related words: act‹1›, approach, assay, consultation‹1›, course‹1›, debate, discussing, discussion‹1›, document, drive‹1›, event‹1›, force‹1›, goal‹1›, incentive, initiative‹1›, investigation‹1›, measure‹1›, potential‹1›, power‹1›, push, reading‹1›, reflection‹2›, report, responsibility‹1›, role‹1›, study, survey‹1›, task, trial, work.
Subsense (ii)
(*Translation:* ζήτημα.)
Synonyms: case‹1›.
Related words: argument, challenge‹1›, concern, dispute, factor‹1›, issue‹2›, judgment, matter, option‹1›, point, problem, question, situation‹1›, sphere‹1›, subject‹1›, theme, topic.
promotion
(*Translation:* προαγωγή, δημιουργία, ανάπτυξη, διάδοση, προώθηση, βελτίωση.)
Hyperonyms: establishing‹1›.
Synonyms: developing‹1›, encouraging, enhancing‹1›, facilitating, fostering, increasing, promoting‹1›, support, transfer‹1›.
Related words: adopting, assisting, beginning, birth, building‹1›, completion‹1›, creation‹1›, demonstration‹1›, deployment‹1›, disseminating, dissemination, distribution, existence‹1›, expansion, flow‹1›, formation, forming, formulating, help, implementation, improving‹1›, increase‹1›, introduction‹1›, launch, launching, manufacture, production, proliferation, propagation, raising‹1›, setting-up, spread, supporting, transmission.
property
(*Translation:* ιδιοκτησία, χαρακτηριστικό, ιδιότητα, ικανότητα, αγαθό.)
Synonyms: capacity‹1›, characteristic, ownership, possession‹1›, possessions, qualification, quality, right‹1›, status‹1›.
Related words: ability, aptitude‹1›, capability‹1›, competence, consignment, definition‹1›, designation, discretion‹1›, duty‹1›, efficiency‹1›, feature, parameter, particular, performance‹1›, power‹1›, principle‹1›, resource‹1›, skill‹1›, supply‹1›, trait.
provision
Hyperonyms: development‹1›.
Subsense (i)
(*Translation:* χορήγηση.)
Synonyms: administering, administration‹1›, allocation‹1›, attribution, award‹1›, awarding‹1›, delivery‹1›, dosing‹2›, giving‹1›, grant‹1›, granting‹1›, issue‹2›, issuing, paying‹1›, therapy, treatment‹1›, use.
Subsense (ii)
(*Translation:* αναφορά.)
Synonyms: clause, communication, complaint, contribution, declaration, document, focus‹1›, identification‹1›, indication, mention, overview, petition, reference‹1›, regard‹1›, report, review, statement.
Subsense (iii)
(*Translation:* καθορισμός.)
Synonyms: defining, definition‹1›, design‹1›, determination‹2›, drafting‹1›, establishing‹1›, formation, forming, formulating, formulation, framing, guidance‹1›, layout, producing, regulation, setting‹1›.
purpose
Subsense (i)
(*Translation:* λόγος, στόχος, σκοπός, αντικείμενο.)
Synonyms: aim, argument, effect‹1›, end‹2›, fact, field‹1›, focus‹1›, goal‹1›, good, ground‹1›, intention, item, mission, need, object, objective, plea‹1›, reason‹1›, subject‹1›, unit, word.
Related words: aspiration, idea, role‹1›, scope‹1›, view, vision‹1›.
Subsense (ii)
(*Translation:* χρήση, λειτουργία.)
Synonyms: management‹1›, operating, task, use, working.
Related words: application‹1›, applying‹1›, conduct‹1›, consumption‹2›, dosing‹2›, effect‹1›, exploitation‹1›, function‹1›, functioning, governance, habit‹1›, implementing, performance‹1›, process, recourse‹1›, running, service, significance, sourcing‹1›, status‹1›, taking‹2›, treatment‹1›, usage‹2›, using, work.
raising
(*Translation:* ανάπτυξη, βελτίωση, αύξηση.)
Hyperonyms: establishing‹1›.
Synonyms: enhancing‹1›, improvement, improving‹1›, increase‹1›, increasing.
Related words: enhancement, expansion, growth‹1›, proliferation, promotion‹1›, rise.
recourse
Subsense (i)
(*Translation:* χρήση.)
Synonyms: use.
Related words: applying‹1›, consumption‹2›, dosing‹2›, exploitation‹1›, management‹1›, operating, purpose, sourcing‹1›, taking‹2›, treatment‹1›, usage‹2›, using, work, working.
Subsense (ii)
(*Translation:* προσφυγή.)
Synonyms: review.
Related words: appeal, petition, proceeding‹1›, remedy‹1›.
reflection
(*Translation:* σκέψη, μελέτη, προβληματισμός, εξέταση.)
Hyperonyms: consideration‹1›.
Synonyms: analysis, assessment‹1›, debate, discussion‹1›, thought.
Related words: assay, considering, consultation‹1›, discussing, document, exercise‹1›, investigation‹1›, problem, project, reading‹1›, report, study, survey‹1›, trial, trouble.
regime

Subsense (i)
(*Translation:* πλαίσιο.)
Synonyms: course‹1›.
Related words: connection‹1›, context‹1›, realm, scope‹1›.
Subsense (ii)
(*Translation:* καθεστώς, σύστημα.)
Synonyms: device‹1›, method, network, pattern, program, programme, scheme, sector, structure‹1›, system‹1›.
Related words: arrangements.
rehabilitation
(*Translation:* θεραπεία.)
Synonyms: therapy.
Related words: care‹1›, treating.
relief
Sense 1
(*Translation:* βοήθεια.)
Synonyms: assisting.
Sense 2
(*Translation:* ανακούφιση.)
Synonyms: relieving.
remedy
Subsense (i)
(*Translation:* παρέμβαση.)
Synonyms: influence, interaction‹1›, interference‹1›, involvement‹1›, response‹3›.
Subsense (ii)
(*Translation:* μέσο, μέτρο, λύση.)
Synonyms: act‹1›, aid‹1›, extent, instrument, legislation, means, measure‹1›, middle, option‹1›, possibility, resort, resource‹1›, rule‹1›, standard, step‹1›, tool‹1›, way.
Related words: answer, approach, decision, determination‹1›, device‹1›, judgment, obligation, reasoning‹1›, response‹3›, result‹1›, settlement‹1›, solution, solving.
Subsense (iii)
(*Translation:* προσφυγή.)
Synonyms: review.
Related words: appeal, petition, proceeding‹1›, recourse‹1›.
remit
Subsense (i)
(*Translation:* καθήκον, αρμοδιότητα.)
Synonyms: function‹1›, mandate, power‹1›, responsibility‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, duty‹1›, expertise‹2›, jurisdiction‹1›, mission, obligation, office‹1›, scope‹1›.
Subsense (ii)
(*Translation:* εντολή.)
Synonyms: order‹1›.
Related words: command‹1›, instruction‹2›.
removal
(*Translation:* απομάκρυνση, αφαίρεση.)
Synonyms: elimination‹1›, expulsion, removing.
resource
Hyperonyms: part.
Subsense (i)
(*Translation:* μέσο.)
Synonyms: aid‹1›, means, middle, remedy‹1›, resort, tool‹1›, way.
Subsense (ii)
(*Translation:* δυνατότητα, ικανότητα.)
Synonyms: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, facility‹1›, power‹1›, qualification, right‹1›, skill‹1›.
Related words: aptitude‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.
response
Sense 2
(*Translation:* ανταπόκριση.)
Synonyms: responding.
Sense 3
Subsense (i)
(*Translation:* παρέμβαση, σχέση, επέμβαση.)
Synonyms: interaction‹1›, interference‹1›, reference‹1›, remedy‹1›.
Related words: association, channel, connection‹1›, influence, involvement‹1›, link, partnership, relation.
Subsense (ii)
(*Translation:* αντιμετώπιση.)
Synonyms: treatment‹1›.
Related words: addressing, countering, handling, management‹1›, meeting‹2›, preventing, processing‹1›, settlement‹1›, tackling, therapy, treating.
Subsense (iii)
(*Translation:* λύση.)
Synonyms: option‹1›.
Related words: answer, approach, decision, determination‹1›, device‹1›, judgment, means, obligation, possibility, reasoning‹1›, remedy‹1›, resort, result‹1›, settlement‹1›, solution, solving, step‹1›.
responsibility
Subsense (i)
(*Translation:* μέριμνα.)
Synonyms: auspice, care‹2›.
Subsense (ii)
(*Translation:* πρωτοβουλία, δράση.)
Synonyms: approach, drive‹1›, effort‹1›, incentive, initiative‹1›, operation.
Related words: action, attempt, effect‹1›, energy‹1›, event‹1›, intervention, measure‹1›, potential‹1›, project, push, role‹1›, scheme, step‹1›, task, venture‹1›, work.
Subsense (iii)
(*Translation:* καθήκον, υποχρέωση, αρμοδιότητα.)
Synonyms: document, duty‹1›, function‹1›, mandate, obligation, power‹1›, remit.
Related words: aptitude‹1›, assignment‹1›, burden‹1›, commitment, competence, condition, debt, expertise‹2›, jurisdiction‹1›, mission, need, office‹1›, right‹1›, scope‹1›.
right
Subsense (i)
(*Translation:* αρχή, ικανότητα, δυνατότητα, αγαθό.)
Synonyms: ability, authority‹2›, capability‹1›, capacity‹1›, competence, consignment, discretion‹1›, facility‹1›, possession‹1›, power‹1›, principle‹1›, property‹2›, qualification, resource‹1›, skill‹1›, supply‹1›.
Related words: agency, aptitude‹1›, association, beginning, body‹1›, commencement, concept‹1›, ease, efficiency‹1›, entity, option‹1›, performance‹1›, policy‹1›, possibility, potential‹1›, quality, rule‹1›, start, status‹1›, term, value, venture‹1›.
Subsense (ii)
(*Translation:* τίτλος.)

Synonyms: certificate, degree, diploma, order‹1›.
Subsense (iii)
(*Translation:* υποχρέωση, αξίωση.)
Synonyms: debt, document, matter, need.
Related words: burden‹1›, claim‹1›, commitment, condition, duty‹1›, idea, obligation, responsibility‹1›.
role
Subsense (i)
(*Translation:* θέση.)
Synonyms: allocation‹1›, attention, job, location‹1›, office‹1›, opinion, place, placing, position, post‹1›, release‹1›, seat, settlement‹1›, site, situation‹1›, spot, status‹1›.
Subsense (ii)
(*Translation:* δράση.)
Synonyms: operation.
Related words: action, approach, attempt, drive‹1›, effect‹1›, effort‹1›, energy‹1›, event‹1›, incentive, initiative‹1›, intervention, measure‹1›, potential‹1›, project, push, responsibility‹1›, step‹1›, task, work.
Subsense (iii)
(*Translation:* σκοπός.)
Synonyms: effect‹1›.
Related words: aim, end‹2›, focus‹1›, goal‹1›, intention, mission, objective, purpose, view.
scope
Subsense (i)
(*Translation:* σημασία.)
Synonyms: attention, extent, implications, importance, interest‹1›, meaning, relevance‹1›, sense‹1›, significance, stress‹1›.
Subsense (ii)
(*Translation:* πλαίσιο.)
Synonyms: course‹1›.
Related words: connection‹1›, context‹1›, realm, regime.
Subsense (iii)
(*Translation:* στόχος.)
Synonyms: focus‹1›.
Related words: aim, aspiration, end‹2›, goal‹1›, idea, intention, mission, need, objective, purpose, vision‹1›.
Subsense (iv)
(*Translation:* αρμοδιότητα.)
Synonyms: power‹1›.
Related words: aptitude‹1›, assignment‹1›, competence, expertise‹2›, function‹1›, jurisdiction‹1›, mandate, remit, responsibility‹1›.
screening
(*Translation:* αναζήτηση, έλεγχος, ανίχνευση, έρευνα, εξέταση.)
Hyperonyms: examination‹1›, event‹1›, consideration‹1›.
Synonyms: exploration, finding‹1›, research, search, seeking, testing.
Related words: detecting, detection, investigation‹1›, recognition.
scrutiny
(*Translation:* εξέταση, έρευνα, έλεγχος.)
Hyperonyms: examination‹1›, event‹1›, consideration‹1›.
service
Hyperonyms: event‹1›, activity‹1›.
Subsense (i)
(*Translation:* υπηρεσία.)
Synonyms: administration‹1›, agency, authority‹2›, body‹1›, centre, department, duty‹1›, employment, facility‹1›, force‹1›, input‹2›, office‹1›, platform, point, sector, unit, work.
Subsense (ii)
(*Translation:* λειτουργία, εργασία.)
Synonyms: task.
Related words: application‹1›, conduct‹1›, effect‹1›, function‹1›, functioning, governance, habit‹1›, implementing, management‹1›, operating, performance‹1›, process, purpose, running, significance, status‹1›, working.
setting
Hyperonyms: establishment‹1›.
Subsense (i)
(*Translation:* κατάρτιση, καθορισμός, διαμόρφωση.)
Synonyms: birth, building‹1›, creation‹1›, defining, definition‹1›, developing‹1›, drafting‹1›, formation, forming, formulating, formulation, framing, introduction‹1›, layout, manufacture, preparation‹1›, preparing, producing, production, provision‹1›, setting-up.
Related words: design‹1›, determination‹2›, establishing‹1›, guidance‹1›, regulation.
Subsense (ii)
(*Translation:* χώρος.)
Synonyms: facility‹1›.
Related words: area, building‹1›, chamber, context‹1›, environment, field‹1›, location‹1›, office‹1›, part, place, region, sector, site, surroundings, terrain.
settlement
Subsense (i)
(*Translation:* θέση, κέντρο.)
Synonyms: centre, foundation‹1›, institution, middle, office‹1›, resort, role‹1›, shop, site.
Related words: allocation‹1›, attention, job, location‹1›, opinion, place, placing, position, post‹1›, release‹1›, seat, situation‹1›, spot, status‹1›.
Subsense (ii)
(*Translation:* αντιμετώπιση.)
Synonyms: treatment‹1›.
Related words: addressing, countering, handling, management‹1›, meeting‹2›, preventing, processing‹1›, response‹3›, tackling, therapy, treating.
Subsense (iii)
(*Translation:* λύση.)
Synonyms: option‹1›.
Related words: answer, approach, decision, determination‹1›, device‹1›, judgment, means, obligation, possibility, reasoning‹1›, remedy‹1›, resort, response‹3›, result‹1›, solution, solving, step‹1›.
sharing
Sense 1
Subsense (i)
(*Translation:* ανταλλαγή.)
Synonyms: exchange‹1›.
Subsense (ii)
(*Translation:* μετάδοση.)
Synonyms: transmitting‹1›.
Subsense (iii)
(*Translation:* κυκλοφορία.)
Synonyms: circulation, flow‹1›, marketing‹1›, mobility‹1›, movement‹1›, traffic‹1›, transit, travel‹1›.
Sense 2
(*Translation:* επιμερισμός, κατανομή, μοίρασμα.)
Synonyms: division‹1›, split, spread.
Related words: allocating, assigning‹1›, distribution.
shift

(*Translation:* μετακίνηση.)
Synonyms: excursion‹1›, journey‹1›, mobility‹1›, move‹1›, travelling.
side
Hyperonyms: area.
Subsense (i)
(*Translation:* τμήμα, μέρος, περιοχή.)
Synonyms: corner, district, field‹1›, land‹1›, location‹1›, place, region, terrain.
Related words: component, destination‹1›, element, limb, moiety, phase, proportion, section, share, spot, strand‹1›, volume‹1›.
Subsense (ii)
(*Translation:* άποψη.)
Synonyms: aspect.
Related words: angle, approach, argument, assumption‹1›, concern, contribution, demand, direction‹1›, idea, message, opinion, option‹1›, perspective, point, position, presumption, statement, view, viewpoint.
skill
Subsense (i)
(*Translation:* γνώση.)
Synonyms: background‹1›, communication, evidence, expertise‹2›, information, insight‹1›, know-how, knowledge, learning, prerequisite, proficiency, study, understanding‹1›.
Subsense (ii)
(*Translation:* ικανότητα, δυνατότητα.)
Synonyms: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, facility‹1›, power‹1›, qualification, resource‹1›, right‹1›.
Related words: aptitude‹1›, ease, efficiency‹1›, option‹1›, performance‹1›, possibility, potential‹1›, property‹2›, quality, status‹1›, venture‹1›.
sourcing
Sense 1
Subsense (i)
(*Translation:* χρήση.)
Synonyms: use.
Related words: applying‹1›, consumption‹2›, dosing‹2›, exploitation‹1›, management‹1›, operating, purpose, recourse‹1›, taking‹2›, treatment‹1›, usage‹2›, using, work, working.
Subsense (ii)
(*Translation:* λήψη.)
Synonyms: administration‹1›.
Related words: administering, grant‹1›, intake‹3›, obtaining‹1›, receipt, receiving, reception‹1›, taking‹2›.
Sense 2
Subsense (i)
(*Translation:* πηγή.)
Synonyms: source‹1›.
Subsense (ii)
(*Translation:* πηγή.)
Synonyms: mechanism.
station
(*Translation:* εργοστάσιο, εγκατάσταση.)
Synonyms: plant‹1›, smelter.
Related words: installation.
status
Hyperonyms: area.
Subsense (i)
(*Translation:* ικανότητα, ιδιότητα.)
Synonyms: capacity‹1›, characteristic, definition‹1›, designation, duty‹1›, property‹2›, qualification, quality.
Related words: ability, aptitude‹1›, capability‹1›, competence, discretion‹1›, efficiency‹1›, performance‹1›, power‹1›, resource‹1›, right‹1›, skill‹1›.
Subsense (ii)
(*Translation:* θέση, κατάσταση.)
Synonyms: context‹1›, position, practice, record‹2›, role‹1›, scenario, situation‹1›, state‹1›.
Related words: allocation‹1›, attention, job, location‹1›, office‹1›, opinion, place, placing, post‹1›, release‹1›, seat, settlement‹1›, site, spot.
Subsense (iii)
(*Translation:* λειτουργία.)
Synonyms: task.
Related words: application‹1›, conduct‹1›, effect‹1›, function‹1›, functioning, governance, habit‹1›, implementing, management‹1›, operating, performance‹1›, process, purpose, running, service, significance, working.
step
(*Translation:* λύση, μέτρο, διαδικασία, ενέργεια, δράση, διάταξη.)
Hyperonyms: development‹1›.
Synonyms: act‹1›, action, approach, attempt, device‹1›, effect‹1›, effort‹1›, energy‹1›, extent, instrument, intervention, legislation, means, measure‹1›, operation, option‹1›, project, remedy‹1›, requirement, rule‹1›, standard.
Related words: answer, clause, configuration, course‹1›, decision, determination‹1›, drive‹1›, event‹1›, exercise‹1›, force‹1›, goal‹1›, ground‹1›, incentive, initiative‹1›, judgment, obligation, order‹1›, possibility, potential‹1›, power‹1›, principle‹1›, procedure‹1›, push, reasoning‹1›, regulation, resort, response‹3›, responsibility‹1›, result‹1›, role‹1›, scheme, settlement‹1›, solution, solving, task, work.
strength
(*Translation:* ισχύς.)
Synonyms: input‹2›, validity‹1›.
subject
Sense 1
Subsense (i)
(*Translation:* ενδιαφέρον.)
Synonyms: importance, interest‹1›, significance, usefulness, value.
Subsense (ii)
(*Translation:* ζήτημα, αντικείμενο, τομέας, θέμα.)
Synonyms: aim, case‹1›, context‹1›, discipline‹1›, factor‹1›, field‹1›, issue‹2›, item, matter, option‹1›, point, problem, purpose, question, sphere‹1›, theme, topic, unit.
Related words: affair‹1›, area, argument, aspect, challenge‹1›, concern, connection‹1›, department, dispute, environment, fact, good, incident, industry‹2›, judgment, object, project, realm, section, sector, situation‹1›.
Sense 3
(*Translation:* ασθενής.)
Synonyms: patient.
survey
(*Translation:* διερεύνηση, έλεγχος, μελέτη, έρευνα, εξέταση.)

Hyperonyms: examination‹1›, event‹1›, consideration‹1›.
Synonyms: assessment‹1›, exploration, exploring, inquiry, investigating‹1›, investigation‹1›, search.
Related words: analysis, assay, consultation‹1›, debate, discussing, discussion‹1›, document, exercise‹1›, project, reading‹1›, reflection‹2›, report, study, trial.
system
(*Translation:* καθεστώς, μέθοδος, σύστημα, μηχανισμός, διαδικασία.)
Hyperonyms: development‹1›.
Synonyms: arrangements, device‹1›, means, measure‹1›, method, network, pattern, procedure‹1›, process, program, programme, regime, scheme, sector, structure‹1›.
Related words: apparatus, approach, equipment, form‹1›, instrument, mechanism, modality, mode‹1›, practice, route, solution, strategy, technique, test‹1›, way.
taking
Sense 1
Subsense (i)
(*Translation:* θέσπιση.)
Synonyms: introduction‹1›.
Related words: adopting, adoption, forming, giving‹1›, introducing, launch, launching.
Subsense (ii)
(*Translation:* διάθεση.)
Synonyms: dissemination.
Related words: allocating, availability‹1›, disclosure, distribution, marketing‹1›, placing, possession‹1›, release‹1›, supply‹1›.
Sense 2
Subsense (i)
(*Translation:* λήψη, απόκτηση.)
Synonyms: administration‹1›, obtaining‹1›, receiving.
Related words: administering, grant‹1›, intake‹3›, receipt, reception‹1›, sourcing‹1›.
Subsense (ii)
(*Translation:* χρήση.)
Synonyms: use.
Related words: applying‹1›, consumption‹2›, dosing‹2›, exploitation‹1›, management‹1›, operating, purpose, recourse‹1›, sourcing‹1›, treatment‹1›, usage‹2›, using, work, working.
Subsense (iii)
(*Translation:* αγωγή.)
Synonyms: treatment‹1›.
Related words: appeal, claim‹1›, petition, plea‹1›, proceeding‹1›, proceedings, therapy.
Subsense (iv)
(*Translation:* διεξαγωγή.)
Synonyms: conduct‹1›.
Related words: deployment‹2›, functioning, performing, progress, running.
test
Hyperonyms: event‹1›, consideration‹1›.
Subsense (i)
(*Translation:* εξέταση, έλεγχος, ανάλυση.)
Synonyms: addressing, analysis, discussing, discussion‹1›, interpretation, overview, review, study, testing.
Subsense (ii)
(*Translation:* μέθοδος.)
Synonyms: procedure‹1›.
Related words: approach, arrangements, form‹1›, means, method, modality, mode‹1›, pattern, practice, process, route, solution, strategy, system‹1›, technique, way.
therapy
Subsense (i)
(*Translation:* αντιμετώπιση, αγωγή, θεραπεία.)
Synonyms: care‹1›, rehabilitation‹1›, treating, treatment‹1›.
Related words: addressing, appeal, claim‹1›, countering, handling, management‹1›, meeting‹2›, petition, plea‹1›, preventing, proceeding‹1›, proceedings, processing‹1›, response‹3›, settlement‹1›, tackling, taking‹2›.
Subsense (ii)
(*Translation:* χορήγηση.)
Synonyms: provision‹1›.
Related words: administering, administration‹1›, allocation‹1›, attribution, award‹1›, awarding‹1›, delivery‹1›, dosing‹2›, giving‹1›, grant‹1›, granting‹1›, issue‹2›, issuing, paying‹1›, treatment‹1›, use.
trade
Sense 1
Subsense (i)
(*Translation:* εμπορία.)
Synonyms: market, trading, trafficking.
Subsense (ii)
(*Translation:* διακίνηση.)
Synonyms: circulation, disseminating, flow‹1›, mobility‹1›, movement‹1›, traffic‹1›, transit, transport‹1›, transportation, travel‹1›.
Sense 2
Subsense (i)
(*Translation:* επάγγελμα, κλάδος.)
Synonyms: industry‹2›, profession‹1›.
Related words: branch, discipline‹1›.
Subsense (ii)
(*Translation:* επάγγελμα.)
Synonyms: occupation‹1›.
traffic
Subsense (i)
(*Translation:* κίνηση.)
Synonyms: move‹1›, organisation‹2›, trend‹1›.
Subsense (ii)
(*Translation:* διακίνηση, κυκλοφορία.)
Synonyms: circulation, disseminating, flow‹1›, marketing‹1›, mobility‹1›, movement‹1›, sharing‹1›, trade‹1›, transit, transport‹1›, transportation, travel‹1›.
transfer
(*Translation:* διάδοση.)
Synonyms: demonstration‹1›, disseminating, dissemination, distribution, expansion, flow‹1›, proliferation, promotion‹1›, propagation, spread, transmission.
transport
Subsense (i)
(*Translation:* διακίνηση.)
Synonyms: circulation, flow‹1›, mobility‹1›, trade‹1›, traffic‹1›, transportation.
Subsense (ii)
(*Translation:* διαβίβαση.)
Synonyms: communication.
Related words: disclosure, transmitting‹1›.
travel
(*Translation:* διακίνηση, κυκλοφορία.)

Synonyms: circulation, flow‹1›, marketing‹1›, mobility‹1›, sharing‹1›, trade‹1›, traffic‹1›, transportation.
treatment
Hyperonyms: consideration‹1›.
Subsense (i)
(*Translation:* αντιμετώπιση, εξέταση, αγωγή.)
Synonyms: addressing, appeal, claim‹1›, countering, handling, management‹1›, meeting‹2›, petition, plea‹1›, preventing, proceeding‹1›, proceedings, processing‹1›, response‹3›, settlement‹1›, tackling, taking‹2›, therapy, treating.
Subsense (ii)
(*Translation:* χορήγηση, χρήση.)
Synonyms: dosing‹2›, provision‹1›, use.
Related words: administering, administration‹1›, allocation‹1›, applying‹1›, attribution, award‹1›, awarding‹1›, consumption‹2›, delivery‹1›, exploitation‹1›, giving‹1›, grant‹1›, granting‹1›, issue‹2›, issuing, management‹1›, operating, paying‹1›, purpose, recourse‹1›, sourcing‹1›, taking‹2›, therapy, usage‹2›, using, work, working.
trend
(*Translation:* κίνηση.)
Synonyms: traffic‹1›.
Related words: move‹1›, organisation‹2›.
usage
(*Translation:* χρήση, εκμετάλλευση.)
Synonyms: exploitation‹1›, operating, use.
Related words: applying‹1›, command‹1›, consumption‹2›, dosing‹2›, exploiting, harnessing, management‹1›, purpose, recourse‹1›, sourcing‹1›, taking‹2›, treatment‹1›, using, utilisation, work, working.
validity
Subsense (i)
(*Translation:* κύρος.)
Synonyms: background‹1›.
Subsense (ii)
(*Translation:* ισχύς.)
Synonyms: strength‹1›.
Related words: input‹2›.
venture
Subsense (i)
(*Translation:* εγχείρημα, επιχείρηση, σχέδιο.)
Synonyms: agency, design‹2›, process, program, programme, undertaking‹1›.
Related words: business‹1›, company‹1›, corporation, enterprise, firm, industry‹2›, start-up, transaction‹1›.
Subsense (ii)
(*Translation:* δυνατότητα.)
Synonyms: facility‹1›.
Related words: ability, capability‹1›, capacity‹1›, competence, discretion‹1›, ease, option‹1›, possibility, potential‹1›, power‹1›, resource‹1›, right‹1›, skill‹1›.
Subsense (iii)
(*Translation:* προσπάθεια, πρωτοβουλία.)
Synonyms: aim, approach, attempt, burden‹1›, drive‹1›, effort‹1›, energy‹1›, momentum‹1›, need.
Related words: incentive, initiative‹1›, responsibility‹1›, scheme.
volume
(*Translation:* μέγεθος, ποσότητα, μέρος.)
Synonyms: amount‹1›, figure, intake‹2›, magnitude, moiety, number‹1›, place, quantity, scale, size.
Related words: component, corner, destination‹1›, element, land‹1›, limb, phase, proportion, section, share, side‹1›, spot, strand‹1›.
work
Hyperonyms: examination‹1›, activity‹1›.
Subsense (i)
(*Translation:* εργασία, δράση, δραστηριότητα.)
Synonyms: business‹1›, capacity‹1›, employment, experience, function‹1›, industry‹2›, instrument, intervention, labour, operating, operation, study.
Related words: action, approach, attempt, drive‹1›, effect‹1›, effort‹1›, energy‹1›, event‹1›, incentive, initiative‹1›, measure‹1›, potential‹1›, project, push, responsibility‹1›, role‹1›, step‹1›, task.
Subsense (ii)
(*Translation:* υπηρεσία.)
Synonyms: service.

TABLE DES MATIERES

Résumé :

Le travail présenté dans cette thèse explore la question de l'acquisition automatique de sens pour la désambiguïsation lexicale dans un cadre de traduction. Partant de l'hypothèse du besoin de conformité des inventaires sémantiques utilisés pour la désambiguïsation dans le cadre d'applications précises, la problématique du repérage des sens se situe dans un cadre bilingue et le traitement s'oriente vers la traduction.

Nous proposons une méthode d'acquisition de sens permettant d'établir des correspondances sémantiques de granularité variable entre les mots de deux langues en relation de traduction. L'induction de sens est effectuée par une combinaison d'informations distributionnelles et traductionnelles extraites d'un corpus bilingue parallèle. La méthode proposée étant à la fois non supervisée et entièrement fondée sur des données, elle est, par conséquent, indépendante de la langue et permet l'élaboration d'inventaires sémantiques relatifs aux domaines représentés dans les corpus traités.

Les résultats de cette méthode sont exploités par une méthode de désambiguïsation lexicale, qui attribue un sens à de nouvelles instances de mots ambigus en contexte, et par une méthode de sélection lexicale, qui propose leur traduction la plus adéquate. On propose finalement une évaluation pondérée des résultats de désambiguïsation et de sélection lexicale, en nous fondant sur l'inventaire construit par la méthode d'acquisition de sens.

Abstract :

This study explores the question of automatic sense acquisition for Word Sense Disambiguation (WSD) in a translation context. On the basis of the need for conformity of the methods and sense inventories used for disambiguation to the requirements of specific applications, the question of sense identification is situated here in a bilingual context and the processing is oriented towards translation.

A sense induction method is proposed which permits the establishment of semantic correspondences of varying granularity between the words of two languages in translation relation. Sense acquisition is done by combining distributional and translation information extracted from a bilingual parallel corpus. Being unsupervised and fully data-driven, the proposed method is language-independent and enables the elaboration of sense inventories relevant to the domains represented in the corpus.

The results of this method are exploited by a WSD method, which assigns a sense to new instances of ambiguous words in context, and by a lexical selection method, which suggests their most adequate translation. Finally, we provide a weighted evaluation of the disambiguation and lexical selection results which relies on the sense inventory built by the sense induction method.