Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle

Mathias Rossignol

N° d'ordre: 3253

THÈSE

présentée

devant l'Université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention INFORMATIQUE

par

Mathias Rossignol

Équipe d'accueil : TEXMEX / IRISA
École doctorale : Matisse
Composante universitaire : IFSIC

Titre de la thèse :

Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle

Soutenue le 26 octobre 2005 devant la commission d'examen

Mme :	Marie-Odile	CORDIER	Présidente
MM. :	Béatrice	DAILLE	Rapporteurs
	Benoît	HABERT	
MM. :	Israël-César	LERMAN	Examinateurs
	Pascale	SÉBILLOT	
	Jean	VÉRONIS	

Remerciements

Le tout jeune docteur chassa encore quelques résidus de coquille, fit quelques pas, et fut confusément conscient de ne pas être arrivé là tout seul…

Je tiens tout d'abord à exprimer toute ma gratitude à Pascale Sébillot, directrice de cette thèse, qui a tout au long de son déroulement fait preuve d'un énergie inépuisable, d'un mélange de rigueur et d'enthousiasme propre à faire avancer le plus réticent des schmilblicks. Grand merci à elle également pour sa patience, sa disponibilité, et son indéfectible confiance — preuve s'il en était besoin que si la foi déplace les montagnes, la confiance les *fait* déplacer.

Je tiens également à remercier Israël-César Lerman, père de la méthode CHAVL, qui a guidé mes premiers pas — et beaucoup des suivants ! — dans le domaine de l'analyse statistique de données, et dont j'ai la faiblesse d'espérer que c'est sans déplaisir qu'il reconnaîtra peut-être, au détour d'une mesure, la trace de ses enseignements.

Je remercie Béatrice Daille et Benoît Habert, rapporteurs de cette thèse, pour le grand honneur qu'ils m'ont fait en acceptant ce rôle, ainsi que pour leurs nombreuses suggestions, qui ont beaucoup apporté à ce rapport tant sur le fond que sur la forme. Plus généralement, je souhaite adresser tous mes remerciements à l'ensemble des membres du jury — donc, outre les personnes déjà citées, Marie-Odile Cordier et Jean Véronis — pour les commentaires, questions et remarques acérées qu'ils ont pu formuler le jour de la soutenance, grâce auxquels je « sors » aujourd'hui de cette problématique avec en vue un horizon plus lointain et plus fascinant que jamais.

Et puisque nos destins varient au gré des influences et coïncidences, il n'est que justice que je remercie ici également le Pr. Michael McKeag, superviseur de mes études de *Bachelor of Science* à la *Queen's University of Belfast*, qui le premier me mit les pieds dans le TAL à l'occasion d'un projet académique de traduction automatique d'Esperanto.

Sur un registre plus personnel, je tiens à remercier pour sa camaraderie Mathieu Lagrange, ami et colocataire au cours de cette période de thèse qui fut aussi la sienne, remue-méninge sans qui les soirées d'hiver eussent été bien longues.

Merci , enfin, à Huyền, comme toujours.

Table des matières

Introduction

Les applications les plus « visibles » du traitement automatique des langues (TAL) au sens large, telles la recherche d'information, la traduction automatique, les systèmes de question-réponse ou le résumé automatique, présentent souvent la caractéristique de requérir pour leur fonctionnement d'importantes quantités de connaissances sur la langue, qu'il s'agisse des outils d'analyse employés (typiquement une forme de modélisation de la langue comme une grammaire formelle) ou des ressources exploitées par ceux-ci (principalement des lexiques morphologiques, sémantiques, *etc.*). Quelle que soit la sophistication de ces applications, la finesse des résultats qu'elles peuvent permettre d'atteindre dépend naturellement de la qualité des connaissances sur lesquelles elles s'appuient, mais aussi de leur pertinence, c'est-à-dire de leur adaptation aux données textuelles traitées. Cette question de pertinence se pose de manière particulièrement cruciale dans le cas des lexiques sémantiques qui, en représentant le sens des mots d'un vocabulaire plus ou moins étendu, doivent donner aux systèmes réalisant une analyse de textes un premier moyen d'accéder au sens de ceux-ci. En effet, les sens des mots employés dans un document sont sujets à d'importantes variations suivant le domaine auquel ce texte se rattache, son auteur, ou l'époque à laquelle il a été écrit ; faute de prendre en compte ces facteurs afin de ne considérer pour chaque mot qu'un ensemble minimal de sens possibles, tous pertinents au regard de la langue employée dans un texte, toute entreprise d'exploitation sémantique de celui-ci semble condamnée à périr sous un flot d'ambiguïtés [Wea55]. Il convient afin d'éviter cet écueil de disposer pour chaque type de texte d'un lexique sémantique approprié, spécialisé, construit de manière à refléter la langue *telle que le texte étudié l'emploie*. Cette exigence soulève une nouvelle difficulté, cette fois d'ordre pratique : les lexiques ainsi définis ne sont que peu ou pas réutilisables, et leur construction manuelle par un ou plusieurs spécialistes est un investissement souvent disproportionné eu égard à l'enjeu de la tâche à accomplir. Ainsi la demande d'une aide ou d'une alternative automatique au travail du lexicographe voit-elle le jour, et c'est dans le cadre de cette problématique que s'inscrivent nos travaux.

Cette préoccupation est l'occasion de l'appropriation par le TAL d'une tendance « phénoménologique » en linguistique, selon laquelle la langue n'est pas un système abstrait, idéal et formalisé, donné *a priori*, et dont les textes

observables ne sont que des réalisations plus ou moins déficientes[1], mais au contraire un phénomène qu'il n'est possible de décrire que de manière empirique, par l'observation des textes. « Le texte travaille [...] la langue » [Bar73], et même, le texte *définit* la langue. La problématique de la construction de lexiques adaptés à des textes particuliers se trouve ainsi reformulée : il ne s'agit plus de détecter les « déviances » de la langue employée dans un texte par rapport à son modèle idéal, mais de tenter d'*apprendre* la langue telle qu'un texte (ou une collection de textes) l'invente. Ces théories du sens que nous nommons phénoménologiques, donnant la primauté à l'observable dans la définition de la langue, fournissent au domaine de la construction automatique de lexiques — et plus généralement, de l'apprentissage de quelque modélisation de la langue que ce soit — les conditions théoriques et méthodologiques de son objectif.

Les conditions techniques en sont aujourd'hui également rassemblées : le développement des technologies de stockage de masse permet la constitution de collections de textes de plus en plus imposantes ; on est ainsi passé en ce qui concerne les ressources de langue anglaise d'un million de mots pour le *Brown Corpus* de 1964 [FK67] aux 100 millions de mots du *British National Corpus* en 1994 [LGB94]. Les capacités de calcul, évoluant dans des proportions tout aussi spectaculaires, rendent pour leur part envisageable l'application de méthodes d'apprentissage automatiques de plus en plus évoluées à ces masses de données. Ces deux facteurs réunis permettent aux chercheurs de donner une réalité informatique à cette nouvelle approche phénoménologique de la langue : à la fin des années 1980 et au cours de la première moitié des années 1990 apparaissent ainsi les premiers travaux visant l'acquisition automatique à partir de corpus de connaissances sur la langue, dans des domaines comme l'apprentissage des catégories grammaticales des mots [HA94], l'inférence de grammaires formelles de la langue naturelle[2], ou l'acquisition d'informations lexicales sémantiques, domaine qui nous concerne ici.

On peut regrouper sous cette appellation des travaux de natures très diverses : extraction de terminologie, et en particulier de termes complexes [Bou94, Dai94, Lau94], découverte de « nouvelles manières » d'exprimer un concept prédéfini [RS97], recherche de mots rapprochés par une relation lexicale donnée (notamment l'hyperonymie [Hea92, JM04]) ou dont les sens, quoique nettement distincts, sont logiquement connectés (par exemple *écrou-visser*) [Cla03]. Les racines de notre travail le rattachent pour leur part à la famille de ceux qui effectuent un rapprochement de synonymes ou quasi synonymes (parfois hyperonymes ou antonymes) afin de former ce que l'on dé-

[1] Cette approche platonicienne est notamment celle retenue par le courant chomskyen de modélisation du langage [Cho57].

[2] Les recherches en inférence grammaticale sont bien antérieures à cette période, mais l'application de cette technique aux langues naturelles était restée très expérimentale [Sik72] ou théorique [Mar80], avant un retour à la fin des années 1980 [Hut88] préludant une apparition régulière dans les conférences consacrées à l'inférence grammaticale dans les années 90 [ICG93].

signe sous le nom de *classes sémantiques*. Les recherches réalisées dans ce domaine trouvent leur fondation méthodologique dans [Gre94], où G. Grefenstette en formalise la problématique et les grands principes : il identifie trois types de proximités entre mots, dites *affinités*, à partir de leur mode d'usage dans les textes, correspondant linguistiquement à des rapprochements syntagmatique, paradigmatique et sémantique. Rejoignant les travaux linguistiques de Z. Harris [Har68, HGR+89], qui postule dès les années 1950 une détermination du sens des mots par leurs usages possibles, G. Grefenstette propose de rassembler les mots de sens proches en les regroupant selon la manière dont ils sont employés dans les textes, autrement dit, de les classer en fonction de la similarité de leurs contextes d'apparition.

Ni la problématique envisagée, ni le principe général des méthodes mises en œuvre n'a évolué de manière significative depuis lors : dix ans après ces travaux fondateurs, la question de la construction de classes sémantiques par comparaison des contextes d'usage des mots étudiés, en employant des méthodes d'apprentissage supervisées ou non, symboliques ou numériques, reste posée. Les principales sources de difficultés rencontrées pour cette tâche sont, d'une part, les possibilités de polysémie des mots, qui entraînent une absence de correspondance directe entre les objets observables (chaînes de caractères) et les objets étudiés (mots-sens), et, d'autre part, la masse des données à considérer, tout à la fois éternellement insuffisantes au regard de la complexité du phénomène à modéliser et trop volumineuses, trop complexes pour les capacités des méthodes d'analyse de données connues.

Ces deux principaux obstacles ont largement conditionné les directions de recherches poursuivies : une première approche fait le choix d'une simplification du phénomène analysé, en se focalisant sur l'étude de langues de spécialités, qui sont à la fois moins complexes (car plus formalisées) et moins sujettes à l'apparition de polysémies (car les sens des mots y sont plus codifiés) ([BHNZ97] par exemple) ; une seconde prend le parti d'une « attaque frontale » exploitant des volumes de textes très importants (plusieurs dizaines, voire centaines, de millions de mots) ([LP01] par exemple). Le travail de recherche est alors principalement consacré au développement d'algorithmes d'analyse adaptés à de telles quantités de données — l'inconvénient de cette approche étant de limiter l'applicabilité des méthodes développées à quelques très gros corpus créés par des spécialistes, s'éloignant ainsi de l'objectif initial d'« à chacun son lexique ». Un point commun à la plupart de ces travaux est en outre de se concentrer exclusivement sur la construction de classes sémantiques, et de mesurer la qualité des données lexicales extraites à la seule aune de la performance atteinte pour cette tâche.

Le travail présenté dans ce document s'inscrit dans une optique plus « exploratoire », puisque nous avons choisi, plutôt que de mettre en œuvre des stratégies de contournement de la complexité du phénomène langagier, de tenter de développer des méthodes permettant d'obtenir des résultats malgré

cette complexité, à partir de données textuelles « ordinaires » et sans faire appel à des informations externes ou à une intervention humaine. En outre, nous ne considérons pas la constitution de classes sémantiques comme une fin en soi : la mise au jour de similarités entre sens de mots n'est en effet qu'une étape pour la construction automatisée de lexiques sémantiques, et l'utilisation sans discernement de ces rapprochements afin de généraliser des énoncés, informations, requêtes, *etc.*, peut générer bien des contresens. Remplaçons « brousse » par « jungle » dans un traité de géographie et nous nous trompons de continent ; « sel » par « poivre » dans une recette de cuisine, nous nous exposons à une surprise culinaire... L'identification des nuances distinguant les significations de mots par ailleurs proches, et précisant ainsi leurs conditions d'usage, est donc une nécessité tant en analyse qu'en synthèse de textes.

Cette constatation nous amène à explorer de nouvelles pistes permettant d'affiner la représentation des sens de mots au-delà de la notion élémentaire de classe sémantique. Alors que les principes de ressemblance de sens entre mots, et de classes reflétant celle-ci, semblent à la fois assez intuitifs et consensuels, tout comme la technique consistant à regrouper pour construire ces classes des mots employés de manière similaire dans une collection de textes, rien d'aussi évident ne s'impose lorsque l'on s'attelle à la tâche de mettre au jour et exprimer les nuances distinctives. C'est pourquoi nous ancrons notre recherche dans les principes d'une théorie linguistique, la *Sémantique interprétative* [Ras96] (que nous appellerons également « sémantique différentielle », en référence au formalisme différentiel de représentation du sens qu'elle emploie et développe), qui nous fournit à la fois le modèle théorique d'une représentation de ces nuances et un guide pour l'exploration des données textuelles à la recherche d'indices révélateurs de celles-ci.

Problématique

L'objectif de notre travail est de rassembler à partir d'un corpus de textes généralistes de taille quelconque (nous travaillons en pratique sur un corpus relativement restreint de 11 millions de mots) des informations concernant les sens des mots employés dans ces textes. L'information sémantique ainsi recueillie est structurée et représentée selon les principes de la sémantique interprétative, théorie développée en France par F. Rastier [Ras96]. Par souci de cohérence et afin d'assurer autant que possible que les connaissances ainsi acquises sont bien celles que nous recherchons en théorie, les méthodes mises en œuvres pour les extraire sont elles aussi inspirées des principes de la sémantique interprétative.

La « philosophie » de notre recherche est de partir à la rencontre des limites du faisable dans le domaine de l'apprentissage totalement automatique sur corpus textuel. Nous nous interdisons donc le recours à des données ou connaissances non inhérentes au texte. Nous évitons également de faire appel à une analyse syntaxique du texte, malgré le fait que les progrès effectués

dans ce domaine permettent aujourd'hui d'envisager une analyse au moins superficielle pour un investissement raisonnable, par souci de ne pas limiter la validité de notre étude à une langue particulière. En effet, si l'intérêt de l'acquisition automatique de lexiques sémantiques est de permettre l'étude de textes pour lesquels aucun lexique adapté n'existe, cela concerne autant les domaines non déjà couverts que les langues pour lesquelles de telles ressources sont rares ou inexistantes, et en particulier les langues dites « minoritaires », pour lesquelles les formalisations syntaxiques sont encore souvent balbutiantes. Quoique n'ayant pu encore tester l'applicabilité de nos travaux à de telles langues, nous les avons menés de manière à ce que rien ne s'y oppose *a priori*. Dans le même ordre d'idées, nous ne nous plaçons pas dans une optique applicative particulière ; l'« expert » auquel il est souvent fait référence en TAL comme instance de décision finale concernant la pertinence des résultats obtenus ne fait donc pas partie de « notre univers », et son intervention relèverait de l'utilisation de nos travaux pour un objectif donné. En conséquence, notre ambition est de pousser les méthodes d'apprentissage automatique au maximum de leur potentiel, et à les adapter aux spécificités des données linguistiques (et non pas au traitement de la langue dans un objectif particulier).

La structure du travail réalisé est guidée par les principes de la théorie linguistique adoptée, que nous détaillons au chapitre 1 (l'ensemble des termes linguistiques que nous employons y voient également leur sens défini de manière plus rigoureuse ; on peut également se reporter pour cela au glossaire, annexe C). Celle-ci suggère un découpage de l'espace sémantique en trois « étages » :

le *domaine* rassemble les mots « parlant de la même chose » — cuisine, relations diplomatiques, course à pied, *etc.* Il constitue ainsi le pendant lexical de ce qu'on reconnaît dans un texte comme l'apparition d'un thème particulier. L'un des points importants de la théorie est qu'au sein d'un domaine, les sens des mots sont stables, autrement dit, il n'existe plus de polysémie (ou du moins pas de polysémie « forte » — nous revenons sur cette nuance au chapitre 1). Cela constitue un apport important pour faciliter l'étude subséquente des textes, et c'est pourquoi, dans un premier temps, nous réalisons parallèlement une structuration du lexique en domaines et un découpage du corpus d'étude en sous-corpus thématiques. Le système de détection de thèmes FAESTOS [3] que nous avons développé à cette fin constitue de manière totalement automatique des classes de mots-clés représentatives des principaux thèmes abordés dans le corpus, sans connaissance *a priori* de ceux-ci. Ces classes sont construites en comparant les profils d'apparition des mots dans les paragraphes du corpus. FAESTOS consiste en une séquence de techniques de classification spécialement adaptée à ces données, de très grandes dimensions mais très peu denses, produisant un résultat dépourvu de « bruit ». Les classes ob-

[3] « Fully Automatic Extraction of Sets of keywords for Topic characterization and Spotting ».

tenues nous permettent de détecter les occurrences des thèmes qu'elles caractérisent, puis d'extraire des textes ainsi sélectionnés le vocabulaire « typique », qui constitue (par définition) le domaine associé ;

le *taxème* correspond à la notion déjà évoquée de classe sémantique, et rassemble des mots qu'il est possible d'employer les uns à la place des autres dans un texte au prix d'une variation de sens mineure ({ *fleuve*, *ruisseau*, *rivière* } peut par exemple constituer un taxème). Nous nous attachons lors de la seconde étape de notre étude à la construction automatique de telles classes à l'intérieur des domaines précédemment définis, et en faisant usage du sous-corpus thématique correspondant. La méthode mise en œuvre consiste à rapprocher les mots employés de manière similaire dans le texte, et se heurte à la faible quantité de données disponibles pour qualifier cet emploi. Les sous-corpus thématiques extraits ne rassemblent en effet que quelques centaines de milliers de mots, ce qui est une taille très restreinte pour un corpus ne relevant pas d'une langue de spécialité. La méthode que nous proposons permet d'obtenir de bons résultats au regard de la faible quantité de données employée, mais ceux-ci sont toujours « bruités » et nécessitent encore un filtrage manuel afin de produire des classes sémantiques exploitables dans la suite ;

les *sèmes spécifiques* précisent, au sein d'un taxème, les nuances distinguant les mots les uns des autres. Ce sont eux qui justifient, alors que les termes d'un taxème sont théoriquement interchangeables en contexte, que l'un d'eux soit choisi de préférence aux autres pour un usage donné (en reprenant l'exemple précédent, *ruisseau* se distingue de *rivière* par sa petite taille). Nous proposons une méthodologie permettant d'automatiser partiellement l'acquisition de ces traits distinctifs, en rapprochant des paires de mots distingués par une nuance similaire (par exemple, *ruisseau-rivière* et *colline-montagne*). Quoique la validation des résultats proposés requière encore une part d'intervention humaine importante, la technique développée permet de mettre au jour une structuration du lexique d'une finesse jusqu'ici inégalée.

En ce qui concerne les deux premiers éléments d'études mentionnés, de nombreux travaux ont d'ores et déjà été menés à bien, que nous présentons à mesure que cela s'avère pertinent dans la progression de ce document. Nous nous en distinguons principalement par notre exigence d'endogénéité totale : les résultats que nous proposons émergent des seules données textuelles étudiées grâce à des outils d'analyse adaptés et ne doivent rien à des données additionnelles. Dans la première étape, FAESTOS réalise de manière totalement autonome la caractérisation et la détection des principaux thèmes du corpus. Dans la deuxième, nous avons approfondi l'étude théorique et statistique du phénomène étudié afin de repousser l'intervention humaine à un rôle de filtrage *a posteriori* de résultats proposés de manière automatique. La troisième étape de ce travail aborde quant à elle une problématique restée jusqu'ici largement inexplorée ; nous en précisons les fondements théoriques et présentons

un premier système d'extraction automatisée de sèmes spécifiques qui, tout prototypique qu'il soit, permet de franchir un seuil qualitatif dans l'acquisition automatique d'informations lexicales sémantiques, et constitue un réel pas en avant dans ce domaine.

Organisation

Le premier chapitre de ce document est consacré à l'approfondissement de notre problématique d'un point de vue linguistique ; nous y abordons en particulier la question de la représentation du sens, et précisons les raisons de notre insistance sur la nécessité d'un ancrage théorique de l'acquisition automatique d'informations lexicales. Nous nous conformons dans les trois chapitres suivants à la séparation en trois parties du travail effectué, tel que suggéré précédemment. Le chapitre 2 détaille donc le système FAESTOS, mis au point afin de découper notre corpus d'étude généraliste en sous-corpus thématiquement cohérents, tout en structurant le lexique acquis en domaines. Nous présentons ensuite au chapitre 3 les techniques élaborées afin de construire, sur chacun de ces sous-corpus, des classes sémantiques rassemblant des ensembles de mots de sens proches, correspondant idéalement aux *taxèmes* de la théorie linguistique. Le chapitre 4, enfin, décrit la méthode développée pour mettre au jour à partir des données textuelles des sèmes spécifiques précisant les différences de sens de mots au sein de ces taxèmes.

Nous nous proposons finalement en conclusion de tirer les leçons de notre travail de recherche, en ce qui concerne à la fois le chemin parcouru, les voies ouvertes à l'exploration et les obstacles inévitables.

Chapitre 1

Lexiques et représentation du sens

La linguistique est une science descriptive, et l'informatique une technologie. Donc, du point de vue épistémologique, l'informatique linguistique est tout simplement une branche de la linguistique appliquée.

F. Rastier, *Sémantique pour l'analyse* [RCA94]

1.1. Introduction

Quelles que soient leur sophistication et leur puissance, les systèmes les plus avancés de simulation physique ou météorologique ne valent que par la pertinence du modèle sur lequel ils fondent leurs calculs. Ce qui est une évidence dans le domaine de la simulation physique semble moins immédiat dans le traitement de la langue, peut-être parce que tout un chacun, disposant d'une connaissance de celle-ci au moins intuitive — ce qui est également vrai de la météorologie — et suffisante dans la plupart des situations quotidiennes où il est nécessaire d'y faire appel — ce qui l'est moins —, peut s'estimer compétent pour l'analyser. Pourtant, des concepts aussi familiers que « sens » ou même « mot » ont fait et font toujours l'objet de nombreux débats ; les racines de leur définition actuelle ne remontent qu'à un peu plus d'un siècle, avec la publication des études de M. Bréal [Bré97] puis F. de Saussure [dS16]. Depuis, les évolutions parallèles de la philosophie et des sciences cognitives, et notamment leur point de rencontre structuraliste, ont fourni à la linguistique le vocabulaire d'une définition systématique des interactions à l'œuvre dans la langue, et les outils permettant d'étudier celle-ci en tant que système, indépendamment de son contexte de production (énonciation) et de sa consommation (compréhension).

Par contraste, il est courant en TAL de se placer d'office dans le cadre d'une application particulière, où ces conditions de production et de consommation

sont justement des facteurs connus. C'est typiquement le cas des systèmes actuels de communication orale homme-machine, mis au point pour répondre à une tâche particulière (réservation de billets d'avion, recherche d'adresses, d'informations touristiques, *etc.*) : plus qu'une analyse des énoncés, le système réalise la mise en correspondance de ceux-ci avec sa connaissance de la situation de communication à laquelle il prend part, par exemple un modèle du locuteur « standard » et de ses intentions supposées — cette approche est particulièrement explicite dans les systèmes fonctionnant par « plans de discours » [Lit85, Car90]. Sans pour autant être aussi directement guidés par la pragmatique, de nombreux travaux de TAL faisant appel à la notion de « sens » définissent celui-ci en fonction d'un rôle communicationnel [Wie48] : le texte est considéré comme un médium ayant pour unique but de transmettre un message concernant un état de fait, à la limite comme un « canal bruité » [Sha48]. Cette vision du texte gouverné par une finalité informationnelle implique de diriger son interprétation selon un besoin d'information défini *a priori* par le cadre applicatif dans lequel prend place l'analyse ; dans le domaine de l'acquisition automatique de connaissances lexicales, ce point de vue trouve son expression sans doute la plus explicite dans l'article d'A. Kilgarriff intitulé *I don't believe in word senses* [Kil97] :

> « *The corpus citations will be clustered into senses according to the purposes of whoever or whatever does the clustering. In the absence of such purposes, word senses do not exist.* »[1]

Si l'on ne peut nier la pertinence pratique de cette approche — tenter d'accéder par apprentissage automatique à un sens du mot « dans l'absolu » n'est ni possible (le choix du corpus étudié étant d'ores et déjà conditionné par une préoccupation particulière) ni même souhaitable (selon notre problématique initiale de construction de lexiques « sur mesure ») —, il semble réducteur de ramener par ce choix l'étude du texte au seul paradigme communicationnel. C'est là, nous semble-t-il, retirer à la problématique sa dimension linguistique et sémantique à proprement parler : d'une part, le texte est un objet en soi, doué d'une certaine permanence qui dépasse l'éphémère d'une situation communicationnelle ; d'autre part, sa production résulte d'une volonté d'*expression*, ce qui ne se réduit pas en général à la transmission d'information ; enfin, d'un point de vue plus lexical, le sens des mots ne peut se résumer à leur mise en correspondance avec le ou les concepts qu'ils dénotent, mais constitue plus largement la brique de base grâce à laquelle peut être tracée la limite entre « ce qui se dit » et « ce qui ne se dit pas » dans une pratique donnée de la langue.

Nous souhaitons pour notre part sinon ignorer totalement l'aspect applicatif de la problématique de construction automatique de lexiques, ce qui ne serait guère réaliste, du moins nous en abstraire suffisamment pour pouvoir

[1] « Les citations du corpus seront regroupées par sens suivant l'intention de celui qui, ou ce qui, réalise le regroupement. En l'absence d'une telle intention, les sens de mots n'existent pas. »

proposer un outil polyvalent réutilisable et spécialisable pour des applications diverses. Nous considérons donc le « sens » comme un objet appartenant au système de la langue indépendamment de tout référentiel, dont la description relève de la sémantique linguistique — seule discipline considérant réellement la langue et le texte, sans faire abstraction des spécificités mentionnées au paragraphe précédent —, et avec lequel il reviendra à des applications concrètes d'interfacer des besoins issus du « monde réel ».

Cette approche constitue la première raison justifiant que nous prêtions une attention particulière à l'ancrage de notre étude dans une théorie linguistique. La seconde tient à l'originalité de la recherche que nous entreprenons ; en effet, la problématique de construction à partir de corpus de classes sémantiques, couramment abordée dans les travaux existants, est presque une évidence étant donné les outils statistiques « classiques » à la portée des spécialistes du TAL : regrouper en classes des objets « ressemblants » selon un certain critère est une préoccupation aussi ancienne que l'analyse statistique informatique de données, et son application directe au regroupement de mots ne requiert pas de formalisme de représentation ou d'interprétation particulier. En revanche, il n'existe pour distinguer de manière fine les sens de mots, comme nous souhaitons le faire, ni mode de représentation, ni technique d'exploration des données textuelles présentant ce caractère d'« évidence » intuitive. Il nous est donc nécessaire de faire appel à un outil théorique proposant à la fois un formalisme de représentation des nuances de sens entre mots, et un guide interprétatif cohérent avec celui-ci pour diriger l'exploration des textes. Le modèle adopté est celui de la sémantique différentielle de F. Rastier, qui a d'ores et déjà fait l'objet de recherches montrant son adaptation aux problématiques rencontrées en TAL [RCA94, Tan97, Beu98, Per04].

Ce chapitre est donc tout à la fois un plaidoyer pour la sémantique linguistique à proprement parler, dont nous tentons de montrer qu'elle est plus que toute autre formalisation (logique, cognitiviste, ontologique) apte à rendre compte des phénomènes signifiants de la langue, et une introduction à ses principes théoriques. Nous y revenons dans un premier temps plus en détail sur le bénéfice, évoqué en introduction, que l'on peut espérer de l'usage d'un lexique sémantique défini à partir du texte constaté, par rapport à une ressource construite manuellement. Pour cela, nous mettons en scène à la section 1.2 les difficultés pratiques pouvant résulter de l'usage du thesaurus WordNet [Fel98], qui constitue aujourd'hui une référence dans le domaine des bases de données sémantiques généralistes construites par des experts. À la section 1.3, nous introduisons de manière générale la problématique de la représentation du sens, ce qui est l'occasion d'un ensemble de mises au point terminologiques précisant le vocabulaire linguistique dont nous pourrons être amené à faire usage. Nous réalisons ensuite un tour d'horizon rapide des diverses familles de formalismes de représentation du sens reconnues à l'heure actuelle (section 1.4), avant d'exposer en détail les principes fondamentaux de la sémantique différentielle, en nous attardant principalement sur les mécanismes

dont nous tirons profit pour notre recherche, section 1.5. Enfin, la section 1.6 nous permet, ayant une meilleure connaissance des principes de cette théorie, de revenir sur celles de ses caractéristiques qui nous paraissent en faire un choix particulièrement pertinent dans le cadre d'un travail d'acquisition automatique d'informations lexicales. Nous discutons pour terminer de la manière dont notre travail se structure selon les lignes de force définies par la sémantique différentielle.

1.2. WordNet

Nous présentons dans un premier temps les principes de la représentation de la signification dans le thesaurus WordNet, qui constitue aujourd'hui, principalement de par sa taille et sa libre disponibilité, la référence en matière de ressource lexicale informatisée. Nous mettons ensuite en avant quelques-unes des limitations intrinsèques à ce type de ressource sémantique « universelle » construite manuellement, ce qui nous permet de préciser les bénéfices potentiels d'une construction de lexique automatique ou semi-automatique basée exclusivement sur les textes.

1.2.1. Description

WordNet [Fel98] est une base de données lexicales électronique dont le travail de construction manuelle, engagé en 1985 par un groupe de linguistes et psycholinguistes de l'université de Princeton sous la direction de G. Miller, est toujours actif aujourd'hui. Il s'agit d'un projet extrêmement ambitieux visant à couvrir l'ensemble de la langue anglaise en précisant pour chaque mot l'ensemble de ses acceptions. La version 2 du système, rendue publique en 2003, rassemble quelque 150 000 mots (dont une grande majorité de noms), soit environ 200 000 paires mot-sens. Cette compilation constitue aujourd'hui une référence en matière de complétude (toute relative, bien entendu), et a inspiré de nombreux travaux visant à aboutir à un résultat similaire pour d'autres langues, comme EuroWordNet pour certaines langues d'Europe de l'Ouest [Vos98] ou BalkaNet pour l'Europe de l'Est [DT00]. Elle est employée dans de nombreux travaux en TAL, comme support d'analyse sémantique ou même d'acquisition lexicale — par exemple afin d'interpréter des termes composés [NV04] ou en tant que « point de départ » à étendre ou spécialiser pour l'adapter à la langue employée dans un texte donné [NV02].

WordNet représente chaque concept par un *synset*, ensemble de termes faisant, dans une de leurs acceptions, référence à une même notion, et donc interchangeables dans un contexte donné (naturellement, un même mot peut appartenir à plusieurs synsets). Ces synsets, qui constituent d'ores et déjà une ébauche d'organisation du vocabulaire, sont placés dans une structure arborescente hiérarchisée par la notion d'hyperonymie. Par exemple, le synset con-

tenant *machine* sera un descendant de celui contenant *artefact*, et un ancêtre de celui rassemblant *hachoir* et *mixeur*. L'organisation du thesaurus, notamment dans les niveaux les plus élevés de l'arbre, reflète une organisation sensément universelle de la langue : ainsi les premiers nœuds séparent-ils des concepts de très grande généralité, tels « action », « état », « abstraction », *etc.* Pour référence, la figure 1.1 reproduit les deux premiers niveaux de la hiérarchie sémantique mise en œuvre dans WordNet pour caractériser les noms.

Les sens des mots sont ainsi représentés par les relations qu'ils entretiennent les uns avec les autres : synonymie par le groupement en synsets, ordonnancement hiérarchique ou « cousinage » par le partage d'ancêtres communs, ainsi que quelques liens transversaux comme l'antonymie ou la méronymie (« est une partie de »), dont l'implémentation est malheureusement encore beaucoup trop parcellaire. WordNet associe en outre à chaque synset une glose permettant d'illustrer le concept représenté. Ce commentaire est destiné en priorité à un utilisateur humain, et semble d'un intérêt limité d'un point de vue informatique. Certaines recherches tentent néanmoins d'exploiter les gloses de WordNet pour la caractérisation des sens de mots (en répondant à la question : « quels mots utilise-t-on pour décrire X ? ») ; ces travaux peuvent également faire usage de définitions issues de dictionnaires papier [LRS04], et l'intérêt des gloses de WordNet dans ce cadre est d'éviter le fastidieux travail de numérisation des textes.

Ayant présenté ses principes de construction, nous pouvons maintenant illustrer les limitations de WordNet « à l'usage » — limitations intrinsèques au modèle de construction manuelle, « experte », adopté et que nous espérons pouvoir dépasser par l'apprentissage basé sur les textes.

1.2.2. Exemples de cas limites

Nous présentons ici sans plus d'éléments d'analyse quelques exemples de catégorisations surprenantes ou peu satisfaisantes en pratique observées dans WordNet ; l'objectif de la section 1.2.3 est d'en fournir une grille de lecture.

Les premiers exemples d'éléments dont le classement peut être jugé discutable apparaissent sur la figure 1.1, sous le synset « *group*, *grouping* » (groupe, regroupement) : tout d'abord, « lanthanides » et « actinoïdes », deux familles d'éléments chimiques, se trouvent descendre directement de ce synset, sans que leur proximité sémantique soit indiquée ni qu'aucun lien puisse être établi entre eux et le synset « élément chimique ». Dans le même ordre d'idées, *sainthood*, dans son acception d'« ensemble des saints »[2], n'est pas classé comme un groupe humain, puisqu'il serait blasphématoire de considérer un saint comme un humain ordinaire. Dans ces deux cas, on constate qu'une extrême précision définitoire limite finalement l'utilité de la représentation du sens développée pour, par exemple, une tâche d'analyse sémantique.

[2] Sens dérivé de son sens premier de « sainteté ».

abstraction
- attribute
- measure, quantity, amount
- relation
- set
- space
- time

act, human action, human activity
- action
- activity
- assumption
- communication, communicating
- distribution
- forfeit, forfeiture, sacrifice
- group action
- hindrance, interference
- inactivity
- judgment, judgement, assessment
- leaning
- legitimation
- motivation, motivating
- nonaccomplishment, nonachievement
- proclamation, promulgation
- production
- rejection
- residency, residence, abidance
- speech act
- stay
- stop, stoppage
- waste, permissive waste
- wear, wearing

entity
- anticipation
- causal agent, cause, causal agency
- enclosure, natural enclosure
- expanse
- location
- object, physical object
- sky
- substance, matter
- thing
- thing

event
- Fall
- group action
- happening, occurrence, natural event
- makeup
- might-have-been
- migration
- miracle
- nonevent
- social event

group, grouping
- actinoid, actinide, actinon
- arrangement
- association
- biological group
- circuit
- citizenry, people
- collection, aggregation, accumulation, assemblage
- community, biotic community
- edition
- electron shell
- ethnic group, ethnos
- halogen
- kingdom
- multitude, masses, mass, hoi polloi, people
- people
- population
- race
- rare earth [element], lanthanoid, lanthanide, lanthanon
- sainthood
- series
- social group
- straggle
- subgroup
- system, scheme
- varna
- world, human race, humanity, humankind (...)

phenomenon
- consequence, effect, outcome, result, event, issue, upshot
- levitation
- luck, fortune
- luck, fortune, chance, hazard
- metempsychosis, rebirth
- natural phenomenon
- process
- pulsation

possession
- assets
- circumstances
- liabilities
- property, belongings, holding, material possession
- transferred property, transferred possession
- treasure
- white elephant

psychological feature
- cognition, knowledge, noesis
- feeling
- motivation, motive, need

state
- action, activity, activeness
- being, beingness, existence
- cleavage
- cognitive state, state of mind
- condition
- condition, status
- conditionality
- conflict
- damnation, eternal damnation
- dead letter, non-issue
- death
- degree, level, stage, point
- dependence, dependance, dependency
- disorder
- dystopia
- employment, employ
- end, destruction, death
- flux, state of flux
- freedom
- grace, saving grace, state of grace
- ground state
- heterozygosity
- homozygosity
- hostility, enmity, antagonism
- illumination
- immaturity, immatureness
- imminence, impendence, forthcomingness (...)
- imperfection, imperfectness
- inaction, inactivity, inactiveness
- integrity, unity, wholeness
- isomerism
- kalemia
- maturity, matureness
- medium
- merchantability
- motion
- motionlessness, stillness
- nonbeing
- obligation
- office, power
- omnipotence
- omniscience
- order
- ornamentation
- ownership
- paternity
- perfection, flawlessness, ne plus ultra
- physiological state, physiological condition
- plurality
- polyvalence, polyvalency, multivalence, multivalency
- readiness, preparedness, preparation
- receivership
- relationship
- relationship
- representation, delegacy, agency
- revocation, annulment
- situation, state of affairs
- skillfulness
- status, position
- temporary state
- tribalism
- turgor
- unemployment
- union, unification
- utilization
- utopia
- wild, natural state, state of nature

FIG. 1.1 – *Deux premiers niveaux de la hiérarchie WordNet*

Nous nous contenterons d'un second exemple illustrant le décalage pouvant exister entre la représentation « rigoureuse » de WordNet et le rôle possible des mots en langue : ancré dans l'imaginaire collectif occidental par une longue tradition et quelques décennies de dessins animés, l'alternative entre les deux animaux domestiques archétypiques chien et chat est sans conteste beaucoup plus structurante que, par exemple, tigre / chat, même si cette dernière proximité est scientifiquement plus établie. WordNet propose pour *chat* la hiérarchie suivante : *chat* < *félin* < *carnivore* < *mammifère placentaire* < *mammifère* < *vertébré* < *chordé*[3] < *animal* < *organisme* < *être vivant* < *objet physique* < *entité*, hiérarchie exacte mais pas forcément pertinente (le premier ancêtre commun à *chien* et *chat* étant « carnivore », et aucune hiérarchie ne proposant le regroupement « animal domestique »). Dans le même ordre d'idées, l'alternative câble / satellite apparaît fréquemment dans les textes traitant des « nouveaux » systèmes de télécommunication ou, plus simplement, de télévision. WordNet propose pour *câble* le sens « télévision par câble », mais n'a pas de synset équivalent pour *satellite*, et le premier ancêtre commun à ces deux mots est « instrument » — catégorie recouvrant aussi bien les moyens de transport que les armes à feu ou articles de toilette.

1.2.3. Interprétation

Le principal défaut imputable à WordNet est la conséquence de l'importante couverture de la langue qui constitue l'un de ses attraits majeurs : la représentation du sens qu'il propose tend à la fois à la généralité et à la « spécialisation universelle ». Il ne peut donc ni prendre le parti d'une représentation correspondant au sens d'une hypothétique « langue de tous les jours », ni assumer le choix d'un « jargon » particulier ; en d'autres termes, les sens des mots proposés dans WordNet ne correspondent pas de manière homogène à une langue unique, réelle et observable. La conséquence de cet entre-deux est que la seule définition acceptable du mot devient la description objective de ce à quoi il fait référence, voire même sa définition scientifique rigoureuse, comme c'est le cas pour l'exemple de « chat » ci-dessus. Plus que d'un lexique, WordNet prend ainsi les caractères d'une ontologie.

Lexique sémantique et ontologie sont deux des principaux paradigmes existant pour la représentation du sens, et correspondent à deux « facettes » de l'information contenue dans les dictionnaires papier classiques :

- le lexique sémantique relève du seul domaine linguistique, et se focalise sur le sens des mots en tant qu'indicateur de la pertinence de leur emploi dans un texte et du rôle qu'ils peuvent jouer dans ce cadre ;
- l'ontologie associe aux mots une description des concepts désignés par ceux-ci, et se place ainsi hors du système de la langue à proprement parler — ou du moins, fait référence à une réalité extérieure à ce système.

[3] Animal possédant un système nerveux organisé autour d'une épine dorsale.

D'un point de vue informatique, l'ontologie relève plus de l'intelligence artificielle que du TAL à proprement parler, et peut parfois être couplée à un formalisme de raisonnement logique.

On peut ainsi dire que le lexique sémantique structure le vocabulaire, tandis que l'ontologie structure le monde[4] par des concepts et fait appel au vocabulaire pour « étiqueter » ceux-ci. Pour sa part, notre travail se focalise exclusivement sur l'aspect lexical du sens, et nous ne retenons que le texte comme objet d'expérience et de validation. Cette approche doit nous permettre d'aboutir à une représentation du sens des mots étroitement adaptée aux textes étudiés, susceptible de prendre en compte des rapprochements inédits entre mots, idiotismes, *etc.*, bref tout ce qui fait la « couleur » d'un usage particulier de la langue, aux associations possiblement inconsistantes et donc de modélisation ontologique problématique. Si nous posons, comme le fait W. O. Quine dès la première page de *Word and Object* [Qui60], l'équivalence entre mot et concept, nous nous limitons à ne pouvoir étudier que les textes suivant les préconisations de la langue canonique qu'il développera dans la suite de ses recherches, ayant constaté l'inadéquation de la langue « ordinaire » à la transmission rigoureuse des idées. Seuls quelques genres textuels (documents techniques, articles juridiques, comptes rendus médicaux, *etc.*) s'approchent de la « langue idéale » définie par Quine, et l'on constate que ces textes sont justement ceux qui se prêtent relativement bien à une analyse sémantique s'appuyant sur des ontologies. Nous ne souhaitons pas imposer dans notre étude de contrainte sur le type de langue étudié, prenant pour objet de notre recherche la langue « ordinaire » dans toute sa richesse.

Repoussant pour ces raisons l'équivalence mot-concept et la représentation ontologique du sens, nous sommes amené à nous interroger sur la manière dont on peut comprendre cette notion de sens sans faire appel à un « monde de référence ». C'est l'objet de la section suivante que de faire un point rapide sur les réponses proposées par la linguistique à cette question. Et puisque la problématique de la lexicographie consiste à étudier le sens des mots, nous nous y attardons également sur la notion de « mot ».

1.3. Du mot au sens : interrogations linguistiques

L'objet d'un lexique sémantique est d'associer à une collection de mots des informations concernant leur sens. Avant de nous interroger sur la manière dont seront représentées et extraites au cours de notre travail ces informations sémantiques, il convient de préciser ce que l'on entend précisément par « mot » et « sens ».

[4] Ou plus rigoureusement « un certain modèle d'un certain monde ».

1.3.1. Quelle unité linguistique pour la lexicographie ?

De manière intuitive, et particulièrement pour un informaticien, le mot est défini comme une suite de caractères séparée de ce qui la suit et la précède par un espace ou une ponctuation… sauf en ce qui concerne les mots composés… et les formules figées… et pas pour toutes les langues… À l'évidence, la définition de la notion de « mot », et plus généralement la question de l'unité linguistique adaptée pour le travail lexicographique, mérite d'être approfondie avant de progresser plus avant dans notre étude.

1.3.1.1. Une brève histoire du mot

Historiquement, le mot est longtemps resté le centre d'intérêt de l'étude linguistique, et a été considéré comme le « point de départ » de la langue : un mot est un signe correspondant à une idée, et l'articulation des mots reflète celle des idées. Au XVIIe siècle, par exemple, « on peut dire en général que les mots sont des sons distincts et articulés dont les hommes ont fait des signes pour marquer ce qui se passe dans leur esprit » [AN83]. Cette équivalence mot-idée s'accommode mal du phénomène de polysémie ; aussi les théoriciens de l'époque s'attachent-ils à effectuer la distinction entre le sens propre, originel d'un mot et ses sens figurés ou dérivés. Cette approche donne une grande importance à l'étude de l'étymologie, notamment au XIXe siècle, et consacre en conséquence dans l'étude académique la primauté du mot écrit sur l'oral. Cette tendance trouve son point culminant avec la publication du *Dictionnaire de la langue française* d'E. Littré (1863-1872).

Dans son *Cours de linguistique générale* [dS16], que l'on s'accorde à considérer comme l'ouvrage fondateur de la linguistique « moderne », F. de Saussure rompt avec la tradition historiciste en affirmant la nécessité d'une étude synchronique*[5] de la langue comme institution humaine observable à un instant donné, indépendamment de son évolution. De ce fait, l'importance donnée à l'écrit disparaît : « langue et écriture sont deux systèmes de signes distincts ; l'unique raison d'être du second est de représenter le premier […] » (*op. cit.*). De Saussure redéfinit du point de vue oral l'entité linguistique élémentaire comme « tranche de sonorité qui est à l'exclusion de ce qui précède et de ce qui suit dans la chaîne parlée le signifiant d'un certain concept » (*op. cit.*), faisant abstraction des conventions de découpage d'un texte en groupes de lettres séparés par des espaces. Il constate que l'unité élémentaire ainsi définie ne correspond plus au mot, qui devient un découpage soit trop fin (dans une locution comme *s'il vous plaît*), soit trop grossier (si dans *désir-eux, malheur-eux,* on distingue la terminaison comme porteuse en propre d'un élément de sens).

Suite à cette remise en question, la définition d'une unité linguistique élémentaire pertinente pour l'analyse a donné lieu au cours du XXe siècle à de

[5] Les mots suivis d'une étoile lors de leur première occurrence sont repris dans le glossaire, annexe C, qui en fournit une définition précise.

nombreux débats. Sans entrer dans le détail de ceux-ci, nous présentons un ensemble de paliers de découpage aujourd'hui relativement consensuels, qui sont notamment ceux retenus dans son travail par F. Rastier.

1.3.1.2. Morphème, lexie, lexème

On nomme *morphème** la plus petite partie indécomposable dotée d'un sens entrant dans la composition des mots (compris au sens du *token* informatique). Ainsi, *innombrables* contient 4 morphèmes : un radical *-nombr-*, deux affixes, et la marque du pluriel. La morphologie est consacrée à l'étude des combinaisons de morphèmes pour former des mots, par opposition à la syntaxe, qui se préoccupe de l'agencement de ceux-ci pour la formation d'énoncés.

Ce qui l'on désigne habituellement par « mot » (termes complexes inclus), correspond à la notion linguistique de *lexie**, groupement de morphèmes « consacré par l'usage » — éventuellement réduit à un unique morphème. Les lexies peuvent être amenées à voir leur forme modifiée par l'ajout de morphèmes flexionnels (dans l'exemple précédent, le *s* du pluriel), qui apportent leur élément de sens à l'ensemble mais ne modifient pas celui de la lexie. L'ensemble des formes fléchies d'une lexie est appelé *lexème**. L'ajout de morphèmes réellement porteurs de sens, par exemple les classiques préfixes « anti- », « pré- », *etc.*, influant par dérivation celui de la lexie, ainsi que la composition de celle-ci avec une autre (conception de mots composés), correspond à la création d'une nouvelle lexie, appartenant à un lexème distinct.

1.3.1.3. En pratique

La plupart des théories du sens considèrent le morphème comme l'échelle de granularité textuelle la plus pertinente pour l'analyse microsémantique*, c'est-à-dire la recherche des composants élémentaires du sens et l'étude de leur combinaison. En revanche, d'un point de vue lexicographique, il est souvent plus utile de disposer de définitions synthétiques de lexies que d'éléments de sens « en kit »… C'est particulièrement le cas en lexicographie informatique, car la composition sémantique des sens des morphèmes pour former celui d'une lexie est loin d'être aussi systématique que l'on pourrait le rêver : *économiste* n'est pas à *économe* ce que *communiste* est à *commun*.

Du point de vue de l'analyse informatique, seul le mot, considéré dans sa définition la plus élémentaire de séquence de caractères séparée du reste du texte par des espaces ou signes de ponctuation, est à la portée immédiate de l'ordinateur. De nombreuses recherches ont néanmoins abouti à la mise au point de systèmes informatiques permettant, d'une part, le regroupement des parties des termes complexes, d'autre part, la « réduction » des mots étudiés à leur forme élémentaire (ou « lemme ») non fléchie (voire même à un radical ou racine). Il est donc courant pour les travaux d'acquisition lexicale de travailler au niveau de la lexie, en faisant abstraction des morphèmes flexionnels.

Maintenant que sont posés quelques principes fondamentaux concernant la définition des *signifiants* étudiés — c'est-à-dire les mots observables —, attardons-nous sur la définition de leur *signifié*.

1.3.2. Référence, signifié, sémème, sémie

La distinction entre référence* (ou parfois « référent ») et signifié* constitue une évolution conceptuelle importante formalisée dans le domaine de la logique par G. Frege à la fin du XIXe siècle [Fre92]. L'objectif de cette distinction est de fonder rigoureusement la possibilité en logique de se prononcer sur la véracité d'un énoncé « dans l'absolu », indépendamment de toute référence à une situation précise (recherche du référent d'un mot dans le monde réel) ou à un individu pensant identifié (recherche du référent dans les objets mentaux de cet individu). Le signifié est défini comme le correspondant du signifiant (symbole, chaîne de caractères ou production vocale) dans un espace sémantique appartenant au même système que celui-ci — système logique formel ou système de la langue.

Dans [Mar67], A. Martinet présente sa conception de la « double articulation » du langage, et précise du même coup ce que l'on peut entendre en sémantique linguistique par la notion de signifié. Le principe de double articulation du langage rend compte du travail de construction, à partir d'une expérience ou d'une pensée individuelle, d'un message sonore compréhensible par l'ensemble des membres d'une communauté linguistique. La première articulation est la décomposition d'un message à transmettre, unique et correspondant à une expérience purement personnelle, en une succession d'unités de faible spécificité partagées par tous les membres d'une communauté. La seconde nous préoccupe moins ici, puisqu'il s'agit de la manière dont les mots eux-mêmes correspondant à ces unités spécifiques communes sont composés à partir d'un ensemble restreint d'unités phonologiques.

A. Martinet désigne par « économie de la langue » la capacité ainsi obtenue d'exprimer une quasi infinité de sens spécifiques à partir d'un minimum d'éléments communs à tous les membres d'une communauté linguistique. On peut penser, pour mesurer cette économie, à l'impossibilité à laquelle on se trouverait confronté de devoir créer et faire accepter « universellement » un cri, un son différent pour chaque message. Ainsi, dans l'exemple de la phrase « j'ai mal à la tête », le type particulier de douleur éprouvée et la désignation d'une tête particulière, celle du locuteur, ne donnent pas lieu à la création de mots spécifiques ; ils ne relèvent pas du système de la langue mais de la compréhension de l'énoncé par son auditeur, autrement dit du domaine de la cognition. La résolution de la *référence**, c'est-à-dire de l'objet du monde réel ou mental désigné par un mot employé dans un énoncé prononcé dans une situation donnée n'est donc pas une préoccupation de la linguistique, ni de la lexicographie. Celles-ci ne considèrent que le *signifié**, portion de sens consensuel concentrée dans une unité linguistique.

Le *sémème**, signifié d'un morphème, et la *sémie**, signifié d'une lexie, peuvent ainsi être considérés comme des points dans ce qui serait l'espace sémantique de la langue. Les différents formalismes de représentation du sens se distinguent principalement selon qu'ils cherchent à décrire ces points par leurs « coordonnées » absolues (représentation descriptive, explicite) ou par leurs positionnements relatifs les uns par rapport aux autres (représentation différentielle).

1.3.3. Sens, signification

F. Rastier désigne la signification comme un « artefact des linguistes » : en effet, si l'on considère le texte comme la seule référence à partir de laquelle la langue se définit, le sens précis du mot varie trop d'un exemple d'usage à l'autre pour qu'il soit possible de mettre en avant un sens « standard ». Et pourtant, les lexiques sémantiques rassemblent bien, sous une forme ou une autre, des descriptions correspondant à un sens « général » ou « standard » du mot, couvrant la majeure partie des sens qu'on lui voit adopter en pratique ou, du moins, permettant que ceux-ci en soient déduits ; c'est ce sens « prototypique » que l'on nomme « signification ». On peut remarquer que cette idée de sens « consensuel » est également centrale dans la théorie d'A. Martinet présentée précédemment.

F. Rastier contourne pour sa part la notion de signification pour ne retenir que celle de sens, lequel ne se définit que par rapport à un contexte d'usage particulier. Dans le cadre du lexique, le sens d'un mot X qui apparaît (jouant le rôle de signification) est celui mis en avant par un contexte « virtuel » constitué de la totalité des mots proches de X dans le lexique. La présentation plus détaillée de la sémantique différentielle à laquelle nous procédons à la section 1.5 nous permet de revenir sur ce principe.

Ayant introduit ces quelques idées fondamentales de la linguistique sémantique et précisé ainsi la nature des objets de notre étude, nous pouvons désormais aborder la question de la représentation du sens, c'est-à-dire du développement d'un formalisme susceptible de décrire au mieux du point de vue de la langue cet évasif objet de notre recherche.

1.4. Représentation du sens

Le plus ancien mode de représentation du sens connu méritant réellement le nom de « formalisation » est celui développé par Aristote, consistant à découper tout ensemble de concepts en deux sous-ensembles rassemblant respectivement ceux possédant et ne possédant pas une certaine caractéristique discriminante : les objets matériels se divisent en vivants et non-vivants, les vivants en mobiles et non-mobiles, *etc.* Le philosophe Porphyre est au IIIe siècle

le premier à exprimer cette structuration sous la forme d'un arbre, posant ainsi l'une des bases incontournables de la pensée occidentale. Naturellement, ce mode de représentation du sens aristotélicien reste fidèle aux principes de Platon, qui refusa à la langue tout autre rôle que celui d'un outil « bijectif » de description de la réalité et des idées. L'arbre aristotélicien est conçu comme une structuration de la réalité, et non pas de la langue puisque celle-ci n'est pas à l'époque considérée comme un système en soi, et est en conséquence adapté à la construction d'ontologies, non de lexiques. Une constatation montrant de manière claire l'inadaptation d'une telle structuration hiérarchique du sens pour la langue est que, dans le cadre d'une pratique linguistique donnée, il existe une séparation assez nette entre « ce que l'on peut dire » et « ce que l'on ne peut pas dire », entre le cohérent et l'absurde : ainsi, s'il est souvent possible de remplacer dans un énoncé *chien* par *chat*, des substituts comme *azote* ou *fourchette* sont hors de question. Il semble donc que se définisse dans le lexique une limite nette entre mots proches et mots clairement distincts, cette dichotomie ne pouvant être reflétée par une théorie instaurant comme l'arbre de Porphyre un « continuum de sens » par le raffinement progressif des définitions des concepts.

Les formalismes de représentation du sens issus de recherches en intelligence artificielle (logiques du premier ordre, floue, modale, *etc.*) ou sciences cognitives (réseaux sémantiques [Qui68], graphes conceptuels [Sow84], *etc.*) relèvent tous de cette approche donnant la primauté au concept (potentiellement un objet mental aux contours absolument flous) sur le mot, qui ne sert qu'à le désigner. Nous intéressant pour notre part au mot, nous ne nous pencherons plus avant que sur les formalismes proprement linguistiques. Nous introduisons ici à titre de référence et de point de comparaison deux théories récentes et reconnues de représentation du sens : la théorie des fonctions lexicales de I. Mel'cuk, et celle du lexique génératif de J. Pustejovsky. Muni de ce léger bagage comparatif, nous présentons ensuite en détail à la section 1.5 les principes de la sémantique interprétative de F. Rastier.

1.4.1. I. Mel'cuk : la théorie Sens-texte

La théorie Sens-texte [MCP95] a été développée par I. Mel'cuk dans les années 1970 ; elle constitue notamment le fondement d'un travail de construction par des experts d'un lexique sémantique décrivant de manière très précise une sélection restreinte de mots du français, le dictionnaire explicatif et combinatoire du français contemporain (DECFC) — les quatre volumes déjà parus de ce dictionnaire, dont la construction est toujours active, rassemblent 510 vocables. Cette théorie fait en particulier usage de fonctions lexicales permettant d'exprimer les relations sémantiques entre mots sous la forme $\mathbf{F}(x) = y$, où **F** est la fonction lexicale, x le mot-clé et y la valeur de la fonction. Par exemple, **Magn**(*pluie*) = *forte* exprime qu'une valeur d'intensification possible de *pluie* est *forte*, **Magn** étant la fonction lexicale d'intensification. Ces fonctions, au

nombre d'une soixantaine pour les plus standards, permettent sensément de représenter dans un formalisme unique un nombre de relations sémantiques suffisant pour préciser le sens de l'ensemble du lexique. Les relations sont dites paradigmatiques quand elles décrivent des relations sémantiques entre mots (comme la généralisation, **Gener**, ou la nominalisation, **S0**), et syntagmatiques quand elles reflètent simplement des relations de cooccurrences (comme les verbes de réalisation associés à un nom, **Real1**, ou l'intensification, **Magn**).

La théorie Sens-texte s'inscrit dans une lignée de recherches théoriques visant à mettre au jour des relations « universelles » en nombre minimal permettant une structuration « complète » du lexique. Du point de vue de l'acquisition automatique de connaissances sémantiques, ces universaux très spécifiques présentent l'inconvénient d'être définis afin de structurer la langue, considérée comme un système abstrait préexistant au texte. Il a néanmoins été montré que certaines des relations proposées par I. Mel'cuk pouvaient être apprise sur corpus grâce à des méthodes semi-supervisées [CL04].

1.4.2. J. Pustejovsky : le Lexique génératif

Les principes exposés par J. Pustejovsky dans [Pus95] peuvent être rapprochés de ceux défendus par I. Mel'cuk en ce sens qu'ils définissent des schémas de relations particuliers entre mots. Ceux-ci sont néanmoins en nombre beaucoup plus faible, quatre structures — naturellement plus souples que les fonctions lexicales — suffisant sensément à fournir la totalité des informations nécessaires pour engendrer grâce à trois mécanismes génératifs les sens en contexte du mot décrit :

- la *structure argumentale* spécifie le nombre, le type sémantique, et la réalisation syntaxique (obligatoire ou facultative) des arguments intervenant dans la définition du mot. Deux arguments sont par exemple nécessaires à la définition du mot « navet » : celui-ci peut en effet être compris comme désignant soit un légume, soit (à un niveau de langue plus familier) une œuvre cinématographique — un seul de ces deux aspects pouvant être pertinent pour une occurrence donnée du mot ;
- la *structure événementielle* recense les événements impliqués dans la sémantique du mot en termes d'aspect (état, processus, transition) et éventuellement de succession chronologique. Elle précisera par exemple que la construction d'une maison prélude à son existence ;
- la *structure des qualia*[6], qui contient la représentation effective de la signification du mot, est elle-même formée de quatre champs, que nous ne détaillons pas ici. Ceux-ci fournissent des prédicats exprimant la manière dont se combinent les événements et arguments décrits précédemment, mettant en avant le « comportement sémantique » du lexème vis-à-vis des concepts qui lui sont liés : un écrou est fait pour être vissé sur un

[6] Du latin « les qualités des êtres ».

boulon, un livre pour être lu, *etc.* ;
– la *structure d'héritage*, enfin, rend compte des relations lexicales entretenues par ce lexème avec d'autres définis par ailleurs.

Peut-être est-il nécessaire d'insister sur le fait que les éléments de description et relations entre entités mentionnés ne sont définis qu'entre entités *lexicales* : conformément à la définition d'un lexique, toute définition est intrinsèque au système de la langue. L'intérêt de cette formalisation réside dans le raffinement qu'elle permet d'atteindre dans la représentation des significations, mais l'on conçoit aisément le revers de cette médaille : la difficulté de construire un lexique de grande taille basé sur ses principes. Il est néanmoins possible de réaliser l'apprentissage automatique d'au moins une partie de cette information, comme cela est montré dans [Cla03], qui développe une méthodologie d'acquisition automatique sur corpus par induction de formules logiques de couples noms-verbes tels que le verbe appartient à la structure des qualia du nom — par exemple, *écrou-visser*. L'association d'un nom avec l'un de ses verbes qualia est en effet fortement susceptible de se trouver réalisée en corpus, et sa recherche revient donc à celle d'un type particulier de cooccurrence, tâche accessible pour l'apprentissage automatique.

Les deux formalismes présentés ont en commun d'être structurés sous forme d'entrées lexicales, et d'imposer pour la construction du contenu de celles-ci un vocabulaire relationnel très précis, un ensemble de « cases » prédéfinies dans lesquelles les informations sémantiques lexicales doivent entrer « de gré ou de force ». Cela s'explique par l'importance que les deux théories donnent à la notion de définition, donnée *a priori* et pré-existant à l'usage en contexte — en particulier pour ce qui est de la théorie Sens-texte : le DECFC a pour objectif explicite de fournir à son utilisateur les informations permettant de générer de manière déterministe toutes les expressions linguistiques correctes de n'importe quelle pensée. Cette approche donnant la primauté à la langue sur le texte implique naturellement que les définitions construites peuvent être difficiles à mettre en correspondance avec les sens constatés en contexte. A. Kilgarriff étudie ainsi dans [Kil01] la capacité des descriptions sémantiques du Lexique génératif à s'adapter aux sens pris par les mots employés de manière « non-standard » dans un texte (par exemple dans le cadre de formulations métaphoriques), et montre que seulement 5 % environ de ces usages correspondent à des sens pouvant être dérivés d'une définition littérale par les règles génératives de J. Pustejovsky.

Par contraste, la sémantique différentielle est beaucoup plus « légère », et se rapporte directement à des phénomènes observables dans les textes eux-mêmes — son optique étant interprétative, et non générative. Nous présentons dans la section suivante ses principes centraux.

1.5. Principes de sémantique différentielle

La sémantique différentielle est la théorie adoptée pour former la base de nos travaux et nous permettre de pousser la représentation du sens automatiquement acquise au-delà de la notion classique de classe sémantique. Nous en exposons ici les principes généraux, tout en mettant en lumière l'intérêt qu'elle présente pour le traitement informatique.

Rappelons que nous employons l'appellation « sémantique différentielle » afin de mettre l'accent sur l'apport principal de cette théorie pour nos travaux, à savoir le paradigme différentiel de représentation du sens qu'elle développe. Dans un cadre plus purement linguistique, l'appellation « sémantique interprétative », qui insiste sur l'importance donnée à la contextualisation des éléments lexicaux et à la définition des opérations d'interprétation, est plus reconnue ; c'est d'ailleurs elle que F. Rastier a retenu pour son ouvrage fondateur [Ras96]. À strictement parler, l'appellation « sémantique différentielle » désigne une famille de théories linguistiques fortement marquées par l'héritage de la linguistique structurale [Pot62] et componentielle [Gre66].

Ce qui suit est tiré en grande partie de deux ouvrages de F. Rastier [Ras96, RCA94], principal responsable du développement de cette théorie en France. Nous présentons dans un premier temps le formalisme de représentation du sens développé dans la sémantique différentielle, qui constitue le centre d'intérêt le plus immédiat de cette théorie au regard de nos travaux. Nous introduisons ensuite les principes d'interprétation ainsi que certains des mécanismes interprétatifs mis en avant par F. Rastier, qui guideront dans la suite notre exploration du texte à la recherche d'indices sémantiques permettant de structurer le lexique.

1.5.1. Sèmes et classes sémantiques

En sémantique différentielle, les sémèmes (signifiés des lexies) sont définis les uns par rapport aux autres par des *sèmes**, dont le rôle est de structurer l'« espace sémantique ». Dans [Ras96], F. Rastier définit le sème comme une « relation binaire entre sémèmes », marquant soit le partage par ceux-ci d'un élément de signification, soit au contraire l'existence d'une différence entre eux. Ce qu'il importe de retenir ici est, d'abord, que le sème en lui-même ne contient rien, ne possède pas de référent. Ensuite, même s'il arrive parfois qu'on le désigne par le terme ambigu de « marqueur sémantique », le sème ne saurait être vu comme un attribut d'un sémème : il ne justifie son existence qu'au regard d'un second sémème, avec lequel il exprime une relation. Enfin, si le sème n'a pas de sens *per se*, il n'en traduit pas moins un rapport sémantique très précis, beaucoup plus porteur d'information que les relations lexicales classiques (synonymie, hyperonymie, *etc.*) ; et c'est bien là ce qui fait la richesse d'une formalisation basée sur la sémantique différentielle par rapport à celle employée dans un thesaurus comme WordNet.

La première conséquence de cela est qu'il est nécessaire de distinguer les sèmes en les nommant ; et bien que ce nommage soit *a priori* arbitraire, la convention est naturellement, même s'il s'agit là d'un glissement risquant de faire oublier la nature réelle des sèmes, de leur donner un nom reflétant ce à quoi ils correspondent pour l'interprétant humain. Par exemple, un même sème rapproche *voiture* de *train* et *vélo* de *autobus*, pour lequel le nom / moyen de transport / semble approprié (les sèmes sont couramment représentés ainsi, parenthésés par une paire de barres obliques). Un autre distingue *rivière* de *ruisseau* aussi bien que *fleuve* de *rivière*, et sera par exemple noté / grand / ; on voit ici combien la notion de *relation* est importante, l'exemple faisant apparaître à l'évidence que l'on ne saurait considérer / grand / comme un simple attribut de *rivière*.

/ moyen de transport / est, dans notre exemple, une relation d'équivalence (pour parler en termes mathématiques, elle est réflexive, symétrique et transitive) qui définit une classe* sémantique : on parle alors de *sème générique**. Ces sèmes sont ceux que l'on pourrait à la limite considérer comme des attributs partagés par les sémèmes qu'ils rapprochent, et qui sont suffisamment généraux pour former une classe sémantique conséquente. On en distingue plusieurs catégories :

– les sèmes *macrogénériques** traduisent des relations de très grande généralité, comme / animé / ou / artefact /. Les classes sémantiques qu'ils génèrent sont appelées *dimensions** ;
– les sèmes *mésogénériques** définissent quant à eux des *domaines**, espaces sémantiques plus restreints et indépendants des dimensions, d'intersection potentiellement non vide, et desquels toute polysémie est bannie — autrement dit, si un même sémème peut être présent dans plusieurs domaines, un domaine ne peut contenir qu'un seul sémème pour un mot donné. / cuisine / peut être classé dans cette catégorie ;
– les sèmes *microgénériques**, enfin, expriment des relations à l'échelle du lexique, du niveau de l'interchangeabilité entre termes dans un contexte donné, et génèrent des classes sémantiques appelées *taxèmes**. Sèmes microgénériques et taxèmes ne sont définis qu'au sein des domaines ; dans le domaine défini par le sème déjà cité / cuisine /, / condiments / est un exemple de sème microgénérique.

Ainsi, pour en revenir au second exemple donné, dans le contexte d'une description de paysage (ce niveau de contexte correspondant à la structuration par domaines), *ruisseau*, *rivière* et *fleuve* seront rassemblés au sein du taxème // cours d'eau // (on note ainsi, entre double-barres obliques, le taxème du sème générique / cours d'eau /). Le sème / grand / donné en exemple précédemment permet donc de structurer ce taxème en précisant les différences sémantiques qui existent entre les mots qu'il contient[7]. Ces sèmes reflétant

[7] Si nous travaillions dans le domaine de l'étude géographique, / grand / ne serait pas le sème spécifique pertinent pour distinguer *fleuve* de *rivière*, mais bien sûr / se jette dans la mer /.

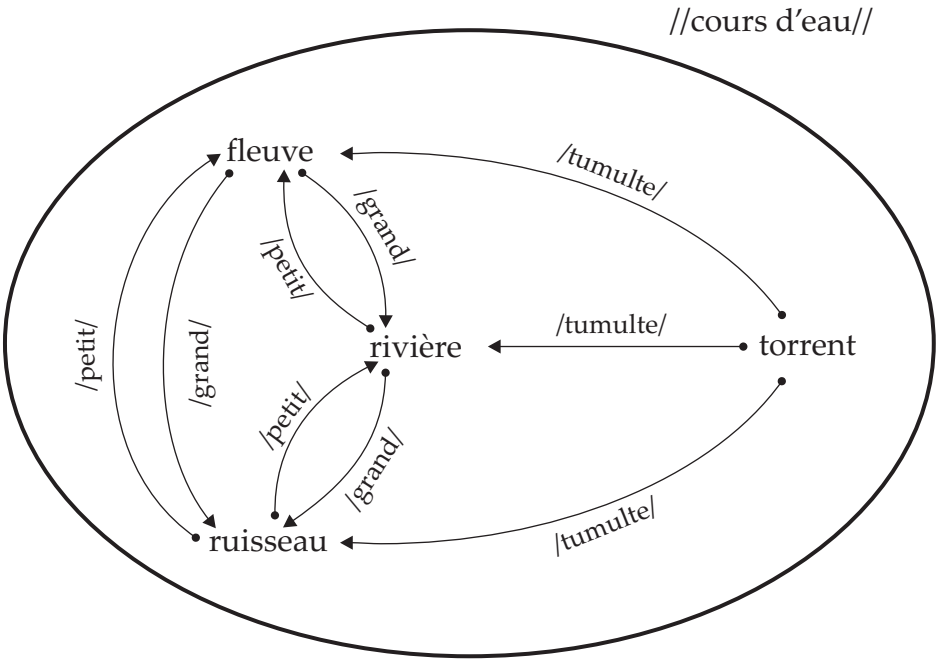


FIG. 1.2 – *Structuration du taxème //cours d'eau// par des sèmes spécifiques. La représentation « A→B » signifie que le sème porté par la flèche distingue A de B*

la spécificité des sémèmes qu'ils distinguent sont nommés *sèmes spécifiques**. Ils sont soit, comme / grand /, des relations antisymétriques, soit des sèmes de trop faible généralité pour que leur manifestation soit justifiée hors du taxème : par exemple, toujours au sein de // cours d'eau //, / tumulte / sera un sème spécifique distinguant *torrent* de la plupart des autres sémèmes (voir figure 1.2) ; on conviendra que la classe des sémèmes caractérisables de « tumultueux » dans le cadre d'une description paysagère est fort restreinte et d'une pertinence lexicale limitée. L'ensemble des sèmes spécifiques d'un sémème est appelé son *sémantème**.

La sémantique interprétative n'a pas pour but une quelconque représentation de connaissances sur les notions référencées par le lexique, et la structuration qu'elle fait de celui-ci n'est pas un réseau sémantique au sens où l'on peut l'entendre dans le domaine des sciences cognitives. F. Rastier place en effet explicitement son travail dans le seul cadre du système fonctionnel de la langue. Dans le même ordre d'idées, le sème ne saurait être considéré comme un prédicat sur les concepts sous-jacents aux sémèmes auxquels il se rapporte : une conséquence de ceci est que même si « possède six pattes » est bien une propriété rapprochant *mouche* de *fourmi*, / six pattes / n'est pas forcément un sème les reliant — selon le domaine, on préférera sans doute / insecte / ou / nuisible /. Cette précision à propos de la distinction entre les propriétés sémantiques d'un mot et les propriétés concrètes de son référent rejoint celle que nous faisions à propos de la distinction entre lexique et ontologie à la section 1.2.3.

Les sèmes n'ont en effet pas vocation à constituer la base d'un raisonnement : il n'est pas question en sémantique interprétative d'évaluer la véracité d'un énoncé, mais d'*interpréter* celui-ci. Nous précisons maintenant la manière dont F. Rastier définit cette notion d'interprétation.

1.5.2. Processus d'interprétation

Comme nous l'avons déjà évoqué, la sémantique différentielle est également appelée sémantique interprétative, en raison de l'importance qu'elle accorde au processus d'interprétation des textes. En effet, si notre présentation a mis en avant l'aspect lexical de la théorie, celui-ci n'est en réalité développé qu'en tant qu'outil — voire même produit dérivé — de l'interprétation des textes, qui constituent le sujet d'étude central de la sémantique interprétative : « Les textes sont l'objet empirique de la linguistique. Isoler des phrases, *a fortiori* des mots, résulte d'une décision méthodologique » [Ras96].

L'interprétation consiste selon F. Rastier à instancier en contexte les « prototypes de sémèmes » définis dans le lexique, par actualisation* et virtualisation* des sèmes prenant part à leurs définitions. Nous présentons ici dans un premier temps les principes de ce mécanisme d'activation / virtualisation avant de préciser comment celui-ci peut permettre de rendre compte de phénomènes parfois considérés comme des manifestations de polysémie. Nous introduisons ensuite la notion d'*isotopie* sémantique*, qui constitue selon la sémantique interprétative le principal guide du processus d'interprétation.

1.5.2.1. Activation et virtualisation de sèmes

Nous avons jusqu'ici considéré le sens comme un objet statique, le sémème étant défini au sein du lexique par les relations qu'il entretient avec ceux qui lui sont proches *en langue*, dans la théorie des regroupements par proximité sémantique. Ce « sens en lexique » est celui que nous avons évoqué précédemment (section 1.3.3) comme jouant en sémantique différentielle le rôle de « signification ». Pour la sémantique interprétative, le sémème n'est considéré dans le cadre du lexique que comme indication de son comportement possible dans le texte. Le sens d'une occurrence du mot est instancié par les mécanismes d'actualisation et de virtualisation de sèmes :

- l'actualisation correspond à l'activation d'un sème pertinent vis-à-vis du contexte du mot : dans l'exemple des cours d'eau, si *rivière* est opposé à *ruisseau*, son sème spécifique / grand / se verra actualisé car il permet d'assurer la définition non ambiguë du mot ;
- la virtualisation est l'opération réciproque, qui permet de « faire taire » les sèmes non pertinents dans un contexte donné : dans l'exemple précédent, le sème / petit / de rivière sera virtualisé[8].

[8]Le terme « inhibition » peut également être employé, mais F. Rastier lui préfère « virtual-

Ces mécanismes guidés par le contexte d'apparition du mot peuvent potentiellement, si ce contexte l'exige, éloigner considérablement le sémème de son sens en lexique. F. Rastier cite ainsi l'exemple de la phrase d'E. Zola : « Guillaume était la femme dans le ménage, l'être faible qui obéit, qui subit les influences de chair et d'esprit » [Zol68]. Dans ce contexte, puisqu'il n'est pas question que Guillaume soit en effet une femme, le sème / féminin / est virtualisé dans *femme* alors qu'il participe, et de manière centrale, à sa définition. En revanche, / faiblesse /, qui n'y participe pas, se trouve actualisé en raison de sa présence répétée dans la suite de la phrase[9]. Correspondant (quoiqu'on puisse en penser par ailleurs) à un lieu commun social et littéraire, l'association femme-faiblesse fait de / faiblesse / un sème dit *afférent* à *femme*, tandis que / féminin /, qui participe à sa définition stricte, en est un sème *inhérent*. Dans le cadre du lexique, les sèmes inhérents sont les seuls pertinents, mais cela peut comme nous le voyons être amendé par le contexte.

Dans l'exemple donné, le sème / faiblesse / a été sélectionné lors de l'interprétation à cause de sa présence répétée dans la phrase : il s'agit de ce que l'on nomme une *isotopie sémantique*, phénomène central dans le processus interprétatif, que nous présentons à la section 1.5.2.3. Nous abordons auparavant rapidement le problème de la polysémie, et la manière dont les mécanismes d'actualisation / virtualisation de sèmes proposés par la sémantique interprétative permettent sinon de le résoudre, du moins de le contourner en partie.

1.5.2.2. À propos de la polysémie

Revenons sur l'affirmation avancée par F. Rastier selon laquelle, au sein d'un domaine, il n'existe plus de polysémie[10]. Cette affirmation ne signifie pas qu'au sein d'un domaine, les mots signifient toujours la même chose en contexte : F. Rastier ne considère comme authentiquement polysémiques que les cas d'homonymie, où un même signifiant possède plusieurs sémèmes nettement distincts (nous avons utilisé à ce sujet dans l'introduction le terme de « polysémie forte »), par exemple « souris », « bleu », « navet », *etc.* Il est néanmoins possible pour un sémème de voir son sens modulé par son contexte d'apparition, ce qui peut prendre l'apparence d'une polysémie si le formalisme employé pour la représentation du sens est très rigoureux (par exemple, une représentation par prédicats logiques). La résolution de l'aspect particulier adopté par un sémème en contexte relève pour la sémantique interprétative de l'interprétation, et non pas du lexique.

---

isation », qui indique plus clairement que le sème, quoique ne prenant pas part au processus interprétatif proprement dit, reste « virtuellement » présent et peut toujours affecter l'impression ressentie par le lecteur à la lecture du texte.

[9] Explicitement dans *faible* et *subit*, plus indirectement dans *obéit* et *influence*.

[10] Nous disons bien « au sein d'un domaine », en tant que niveau de structuration du lexique, et non pas au sein d'un texte évoquant le thème correspondant à ce domaine.

F. Rastier cite dans [Ras96] l'exemple de l'ambiguïté polysémique apparemment insoluble « il refait son cours tous les ans ». Le professeur dont il est question est-il particulièrement zélé, amendant chaque année le contenu de son enseignement, ou répète-t-il année après année le même cours identique ? Seule la connaissance d'un contexte d'énonciation plus large pourrait permettre de répondre à cette question, mais quelle qu'elle soit, le sémème *cours* présent est le même : c'est par le processus d'interprétation que prendra le dessus son aspect de « contenu » ou de « spectacle ». F. Rastier note encore, toujours dans [Ras96] : « une bouteille cassée est-elle encore une bouteille ? Nous dirions que dans *bouteille cassée* le contexte *cassé* virtualise le trait / récipient / dans *bouteille* ».

Cette précision effectuée, revenons au mécanisme d'interprétation, et en particulier à sa composante la plus importante, la notion d'isotopie.

1.5.2.3. Isotopies sémantiques

Le processus de sélection / affectation de sèmes est en particulier dirigé par le principe d'*isotopie sémantique*, principe introduit par A. J. Greimas dans [Gre66] et largement repris et développé par F. Rastier, selon lequel l'impression de cohérence à la lecture d'un texte est due à la répétition de sèmes ou groupes de sèmes, à toutes les échelles de granularité sémantique — on pourrait presque parler de « principe de redondance ». Ainsi, si « le chien remue la queue » semble plus naturel que « le chien remue la cloche », c'est parce que dans le premier cas l'énoncé fait apparaître à deux reprises le sème / animal / , dans / chien / et / queue /. Selon un principe similaire, si lors d'un test de perception le sujet annonce : « la première forme est bleue, la seconde verte, la troisième rouge et la quatrième triangulaire », sa réponse pourra être jugée à la fois totalement valide et manquant cruellement de pertinence — le principe d'isotopie appelle le sème / couleur / pour la dernière réponse.

L'isotopie peut se manifester à différentes échelles dans le texte, et à plusieurs niveaux sémantiques. Ainsi, F. Rastier traduit sous la forme d'isotopie les principes d'accord grammatical : isotopie d'un sème de nombre, genre, *etc.* À l'échelle des groupes syntaxiques simples, l'exigence d'isotopie peut conduire le processus interprétatif à réaliser une *propagation de sème* par laquelle un sémème se trouve porteur d'un sème qu'il ne possède pas en lexique. Il peut s'agir d'un sème afférent, comme dans l'exemple de la phrase de Zola citée à la section 1.5.2.1, ou d'un sème totalement étranger, comme cela peut souvent se produire en langage imagé : dans « le ruisseau gazouille », *ruisseau* se trouve par la force du contexte porteur du sème / animal / (ou plus précisément, suivant le contexte éventuel, / oiseau / ou / bébé /).

L'isotopie s'observe de manière particulièrement évidente au niveau mésogénérique (domanial) et à une échelle textuelle relativement grande, introduisant une « coloration thématique » dans un segment de texte. Nous avons

« En fait, les *réseaux électroniques* sont majoritairement le **domaine réservé** des **élites** et de leurs ***satellites*** - principalement les **classes moyennes** des pays du Sud, en Corée, en Inde ou au Brésil. Pour les **privilégiés**, *naviguer* dans les *réseaux* est le signe d'un *progrès* capital. Mais considérons les faits : en 1995, le nombre d'*ordinateurs* personnels (*PC*) en usage dans le monde était d'environ 180 millions , pour une population globale de presque six milliards d'individus. La possibilité d'accéder au *réseau* était donc limitée à 3 % de personnes. En 1995, un petit nombre de pays **riches**, représentant environ 15 % de la population mondiale, **possédaient** environ les trois quarts des principales *lignes téléphoniques*, sans lesquelles, par le biais d'un *modem* (*modulateur-démodulateur*) on ne peut accéder à *Internet*... Plus de la moitié de la planète ne s'était jamais servie d'un *téléphone* : dans quarante-sept pays, il n'y avait même pas une seule *ligne* pour cent habitants ; or, d'après l'Union internationale des *télécommunications*, on peut affirmer qu'il dessert les **besoins** d'un pays lorsque nul n'est situé à plus de cinq kilomètres d'un *téléphone*. »

FIG. 1.3 – *Paragraphe évoquant le thème des nouvelles technologies de l'information et de la communication extrait du corpus du* Monde diplomatique *(Dan Schiller,* Les marchands du cyberespace*, mai 1996, pp. 15-20), sur lequel nous avons fait apparaître les mots participant aux trois principales isotopies domaniales qu'il développe : sur fond gris, «* / monde réel / + / population / + / pays / *» constitue l'ossature du discours ; l'auteur projette sur ce canevas «* / communication / + / distance / + / électronique / *» (en italique), et fait ressortir par cette confrontation les déterminants économiques à l'accès aux nouvelles technologies — «* / richesse /, / classe / *», en gras*

choisi afin d'illustrer ce phénomène de sélectionner aléatoirement un paragraphe dans le corpus employé pour notre recherche (il s'agit d'une collection d'archives du mensuel *Le Monde diplomatique*, rassemblant quelque 11 millions de mots), et de mettre en évidence les isotopies mésogénériques apparaissant dans ce texte. Le résultat de cette expérience est donné à la figure 1.3 : on peut constater que le processus interprétatif guidé par la triple isotopie détectée nous permet de mettre au jour l'intention argumentative du texte.

Nous ne prétendons pas épuiser dans cette présentation l'intégralité des précisions apportées par F. Rastier concernant les processus d'interprétation. Nous somme amené au cours de ce manuscrit à apporter à ces « fondamentaux » divers compléments, à mesure qu'ils s'avèrent pertinents pour notre travail. Pour l'heure, la section suivante est consacrée à une rapide synthèse soulignant l'adaptation de la sémantique interprétative à la tâche d'acquisition sur corpus d'informations lexicales.

1.6. Sémantique différentielle et acquisition automatique d'informations lexicales sémantiques

Les deux termes que nous employons pour désigner la théorie linguistique de F. Rastier, « sémantique différentielle » et « sémantique interprétative », font référence aux deux facettes principales de son propos : une représentation du sens dans le lexique (nous dirions « signification » si F. Rastier n'écartait pas explicitement cette notion du cœur de sa théorie) fondée sur le paradigme différentiel, et une manifestation de ce sens en contexte rendue accessible à l'analyse rationnelle grâce à l'explicitation du processus d'interprétation. Ces deux aspects sont autant d'atouts de cette théorie pour son utilisation dans un cadre d'acquisition automatique d'informations lexicales : ils nous permettent en effet de disposer d'un formalisme de représentation du sens adapté à l'acquisition automatique, ainsi que d'un guide de lecture pour l'analyse des données textuelles.

1.6.1. Sémantique différentielle

L'un des caractères distinctifs les plus marquants de la sémantique différentielle dans la version de F. Rastier par rapport aux théories linguistiques du sens telles que celles que nous avons pu présenter à la section 1.4 est son absence totale de recours à un métalangage de description linguistique : aucun vocabulaire ni aucune structure n'est employé sinon ceux issus de la langue elle-même. Bien entendu, les sèmes sont nommés, mais ces noms sont arbitraires, et il ne leur est demandé d'avoir un sens que parce la présentation qui en est faite s'adresse à un lecteur humain. Formellement, domaines et taxèmes peuvent être complètement définis par la simple donnée des sémèmes qu'ils rassemblent, caractérisant du même coup de manière suffisante le sème méso- ou microgénérique qui les sous-tend. Les sèmes spécifiques, quant à eux, peuvent être exprimés par la donnée des paires de sémèmes qu'ils distinguent. La sémantique différentielle de F. Rastier nous fournit ainsi le moyen d'une représentation du sens réellement intrinsèque à la langue. N'exprimant que des contraintes minimales concernant la forme des informations sémantiques, elle ne limite pas *a priori* leur contenu, et est ainsi apte à capturer toute la richesse des relations qu'un apprentissage automatique à partir de corpus peut faire émerger.

De plus, si les domaines et taxèmes correspondent à des niveaux d'analyse déjà largement explorés, par exemple par les nombreux travaux existants en extraction automatique de classes sémantiques, les sèmes spécifiques constituent grâce à leur combinaison de simplicité formelle et de grande puissance expressive un outil permettant d'espérer atteindre un nouveau niveau de détail dans la représentation du sens.

1.6.2. Sémantique interprétative

Cette représentation du sens trouve en outre sa justification dans les textes eux-mêmes, dans l'usage constaté des mots tel qu'il se prête au processus d'interprétation. F. Rastier insiste tout particulièrement dans [Ras96] sur son objectif de description rationnelle du processus interprétatif, et réalise une typologie détaillée des isotopies, phénomènes d'inhibition, activation ou propagation de sèmes, *etc.*, qui sont selon lui les « mouvements de base » de cette interprétation. Cette étude fournit à l'informaticien linguiste un guide inestimable pour diriger l'analyse automatique du texte à la recherche de motifs, relations, ou « affinités », pour reprendre le terme de G. Grefenstette, permettant d'induire une structuration du lexique.

Toute adaptée que la théorie soit à nos objectifs, la traduction de ses principes linguistiques en décisions implémentatoires informatiques n'est pas une tâche triviale, et les choix effectués à ce niveau conditionnent l'ensemble du travail réalisé.

1.6.3. Exploitation informatique

Comme nous l'avons vu, la sémantique différentielle propose une hiérarchie du sens dans le lexique organisée en trois « étages » : le domaine est une structuration du lexique à l'échelle du discours, et rassemble des mots susceptibles d'utilisation conjointe pour évoquer un thème donné ; le taxème rassemble des mots de sens proches, échangeables en contexte ; enfin, au sein de ce dernier, les sèmes spécifiques marquent les distinctions de sens du niveau de la sélection du « mot approprié » dans un contexte donné.

Notre recherche adopte ces trois niveaux comme étapes successives de la description du sens. Dans un premier temps, nous entreprenons donc de découper notre corpus d'étude en sous-corpus thématiques, au sein desquels se développe de manière privilégiée le discours lié à un domaine donné, employant les mots de ce domaine. Cette première étape nous permet de « stabiliser » les sens de mots dans les sous-corpus, et ainsi de contourner la nécessité, souvent évoquée en acquisition automatique d'informations lexicales, de travailler sur un corpus spécialisé. Nous entreprenons ensuite de rassembler les mots des domaines ainsi identifiés en taxèmes — assimilables aux classes sémantiques dont la construction constitue une préoccupation centrale dans les travaux existants d'acquisition de ressources sémantiques. Enfin, nous mettons en évidence, au sein de ces taxèmes, des sèmes spécifiques exprimant les différentiations sémantiques existant entre les termes rassemblés.

Cette procédure en trois étapes, directement guidée par les principes de la sémantique différentielle, nous permet d'espérer pouvoir construire à partir des seules données textuelles rassemblées dans un corpus des lexiques sémantiques reflétant les sens des mots tels qu'ils sont en effet employés dans les

textes. Si ce guidage théorique peut constituer un atout précieux pour réaliser une exploitation rigoureuse des textes au cours des deux premières étapes, son apport majeur est de nous permettre de pousser par la troisième étape le niveau de granularité de la représentation sémantique acquise jusqu'à un degré de finesse de distinction de sens jamais encore atteint par d'autres travaux d'acquisition automatique d'informations lexicales.

Pour chacune des trois étapes mentionnées, l'analyse théorique des phénomènes linguistiques considérés est accompagnée d'un investissement systématique de tous les outils et ressources mathématiques, statistiques et algorithmiques à notre portée, afin de mener aussi loin que possible l'automatisation des tâches entreprises. Cette approche nous permet en particulier de proposer pour la première étape de notre recherche un système totalement autonome de caractérisation et détection de thèmes, décrit au chapitre suivant.

Chapitre 2

Domaines et corpus thématiques

Il suffit d'ajouter « militaire » à un mot pour lui faire perdre sa signification. Ainsi la justice militaire n'est pas la justice, la musique militaire n'est pas la musique.

G. Clémenceau

2.1. Introduction

Nous présentons dans ce chapitre la première partie de notre travail d'acquisition d'informations lexicales, telle que suggérée par la hiérarchie domaine / taxème / sémantème de la sémantique différentielle. Notre objectif est double : nous souhaitons d'une part regrouper le vocabulaire employé dans un corpus en classes mésogénériques, ou domaines, et d'autre part découper ce corpus d'étude initial en sous-corpus thématiques au sein desquels nous pourrons étudier l'emploi des mots en partant d'une hypothèse de cohérence domaniale. Comme nous l'avons précisé au chapitre précédent, les domaines constituent des classes sémantiques d'assez grande généralité, dont l'intersection est potentiellement (et même couramment) non vide. Formellement, ils sont définis comme rassemblant l'ensemble des mots partageant un même sème mésogénérique ; en pratique, on constate que ces sèmes rapprochent des mots « que l'on emploie pour parler d'un thème donné ». Thème et domaine peuvent ainsi être considérés comme des notions « duales » dans les deux espaces du texte et du lexique, respectivement. Le texte constituant pour nous la seule donnée observable, nous choisissons d'aborder le problème des domaines sous l'angle du thème, définissant chaque domaine par la donnée des textes évoquant le thème correspondant, dont il « suffit » d'extraire le vocabulaire spécifique. Nous nous concentrons donc dans ce chapitre sur le travail de détection et caractérisation de thèmes au sein d'un corpus généraliste, et revenons en conclusion sur les méthodes simples qu'il est possible de mettre en œuvre pour déduire de ce découpage une organisation en domaines du vocabulaire qu'il emploie, réalisant ainsi une première structuration du lexique.

On recouvre couramment sous la dénomination « détection de thèmes » une part plus ou moins importante, suivant le travail considéré, de trois problématiques distinctes (quoiqu'il soit impossible de les séparer totalement en pratique) : caractérisation des thèmes abordés dans un texte, segmentation de ce texte en sections thématiquement homogènes, et détection (ou repérage) des thèmes à proprement parler.

Caractérisation : il s'agit de réaliser un inventaire et une description des thèmes supposément abordés dans le corpus. Cette partie du travail est naturellement ignorée par les nombreux travaux réalisant la détection de thèmes posés *a priori*. Fidèle autant aux principes de la sémantique interprétative qu'au propos d'automatisation totale de notre travail de recherche, nous considérons qu'il est artificiel de prédéfinir une collection de thèmes « standard » à plaquer sur la réalité du texte, mais qu'il convient de faire émerger la définition des thèmes des données textuelles elles-mêmes. Nous réalisons donc de manière automatique l'apprentissage de descriptions caractéristiques des thèmes abordés dans le corpus étudié.

Segmentation : cette opération consiste à identifier dans un texte les points de changement de thème, et de le découper ainsi en segments thématiquement homogènes. Elle peut être considérée indépendamment de toute préoccupation de détection, dans le but par exemple de faire apparaître la structure d'un document, ou de manière conjointe avec elle afin de cibler de la manière la plus « serrée » possible les apparitions des thèmes. Pour notre part, ayant pu constater la fréquence des phénomènes d'entrelacement de thèmes multiples (comme c'est le cas sur la figure 1.3, page 34), nous avons choisi d'ignorer cette étape et de définir *a priori* le paragraphe comme unité textuelle « atomique » pour l'étude des apparitions de thèmes. On peut en effet constater sur le texte de la figure 1.3 que le paragraphe reproduit semble bien constituer une échelle « naturelle » de développement du thème. Si la validité de cette hypothèse de cohérence des paragraphes n'est pas garantie de manière générale, elle se vérifie bien pour la plupart des corpus écrits « classiques » (journalistiques, scientifiques, juridiques, *etc.*), qui résultent d'une pratique professionnelle de la langue privilégiant la clarté de la structure du texte. Ayant ainsi fait le choix d'un niveau de structuration textuelle d'une taille relativement importante, nous admettons en revanche qu'un segment (paragraphe) puisse aborder simultanément plusieurs thèmes, et que l'évocation d'un même thème puisse s'étendre sur plusieurs paragraphes successifs — le système présenté est tout à fait indifférent à ces deux facteurs.

Détection : la détection de thèmes est l'opération consistant à mettre en correspondance les thèmes et segments de textes définis ou extraits par les deux tâches précédentes. À cette fin, il est fait usage d'un critère permettant de reconnaître les segments de textes abordant chacun des thèmes

définis, par exemple en associant à chacun de ces thèmes un modèle de la langue employée lorsqu'il est abordé, et en mesurant l'adéquation des segments de textes avec ces modèles. Nous avons pour notre part fait le choix d'un formalisme de caractérisation des thèmes également adapté au rôle de critère de détection : il s'agit de classes de mots-clés thématiques, que nous extrayons du corpus de manière automatique par analyse statistique des répartitions de ses mots dans ses paragraphes ; nous considérons que l'occurrence conjointe de plusieurs mots-clés d'une même classe dans un paragraphe y révèle la présence du thème caractérisé par la classe.

Au-delà de notre intérêt direct pour l'acquisition d'informations lexicales, la connaissance du thème abordé dans un extrait de texte peut constituer une information précieuse pour la mise en œuvre de nombreuses applications du traitement automatique des langues : c'est par exemple le cas en filtrage ou catégorisation de textes, en recherche d'information [MS05] où l'on peut souhaiter centrer l'analyse sur des portions de textes reconnues comme abordant un thème lié au sujet traité, ou pour le résumé automatique, où la connaissance de la structuration thématique du texte peut être une source d'information capitale [BN00]. C'est une préoccupation que l'on retrouve aussi dans des domaines connexes, comme la reconnaissance de parole [BdMEBS97, BLB01]. Nous considérons donc également la solution présentée dans ce chapitre comme un système à part entière, répondant à une problématique intéressante en tant que telle, et que nous avons baptisé FAESTOS. Par analyse statistique des répartitions des mots d'un corpus étudié sur ses paragraphes, FAESTOS génère de manière totalement automatique, sans requérir d'intervention humaine, de préconnaissance des textes (notamment en ce qui concerne le nombre et la nature des thèmes développés dans le corpus), ou de source de données extérieure, des classes de mots-clés caractéristiques des principaux thèmes abordés dans un corpus. Ces classes nous permettent, selon un critère de coprésence de mots-clés dans un même paragraphe, de détecter les occurrences de ces thèmes, mais aussi de les « nommer » grâce à un triplet de mot sélectionnés automatiquement parmi ceux qu'elles rassemblent. L'accent placé sur la tâche de caractérisation, souvent secondaire dans les travaux portant sur l'analyse thématique de textes, s'explique du point de vue de notre tâche globale de construction de lexiques : il est naturellement préférable pour l'exploitabilité de ceux-ci que les domaines en lesquels leur vocabulaire est structuré soient désignés de manière explicite et aisément compréhensible.

La section 2.4 est consacrée à une étude plus approfondie des principes de sémantique interprétative sur lesquels se fonde le système FAESTOS, ainsi qu'à une description générale de son fonctionnement. Les détails de sa mise en œuvre pratique font l'objet des sections 2.5 à 2.7. Nous procédons ensuite à une évaluation de la performance du système, section 2.8, avant de revenir à notre problématique initiale d'extraction d'informations lexicales sémantiques.

Au préalable, nous présentons dans la section suivante quelques travaux qui abordent de diverses manières la problématique de détection de thèmes, et constituent un référentiel permettant de préciser notre approche, puis apportons en section 2.3 quelques éléments d'information concernant l'entour de notre recherche : le corpus que nous employons pour réaliser nos expériences, ainsi que le principe général des méthodes statistiques dont nous sommes amené à faire usage.

2.2. Quelques autres approches de la détection de thèmes

La détection de thèmes au sens large, c'est-à-dire les trois sous-problématiques dont nous avons montré qu'elle se composait, a donné lieu à de nombreuses recherches tirant profit de sources d'informations diverses : indices linguistiques révélateurs d'un changement ou d'une annonce de thème [LP95, FGMP01, BHDM01], indices extralinguistiques tels la prosodie dans le cas de corpus oraux [THTSS01] ou la structure d'hyperliens et de co-citation pour les pages Internet [HDZS01], ou enfin cohésion lexicale (ou la rupture de cette cohésion) [Hea94, SSBM96, FG01] — principe selon lequel deux segments de textes « parlant d'une même chose » sont susceptibles d'employer pour ce faire les mêmes mots, et réciproquement. Ces derniers travaux, les seuls à ne s'appuyer que sur les données textuelles brutes sans apport de pré-connaissances linguistiques, sont les plus proches de nos préoccupations. Nous présentons ici les principes de quelques-uns des plus marquants d'entre eux, ce qui nous permet de préciser nos choix en ce qui concerne tant les méthodes mises en œuvre que les objectifs poursuivis. Cette présentation ne s'étend pas aux travaux portant sur la classification de documents [LHKK96], domaine connexe ayant donné lieu à une littérature fournie mais d'un intérêt direct limité dans le cadre de notre travail.

TEXTTILING [Hea94] réalise le découpage d'un texte en groupes de paragraphes successifs portant sur un même thème en se basant sur une mesure de similarité lexicale entre séquences consécutives de mots. Tous les 20 mots, l'algorithme calcule la « ressemblance » entre les listes des 100 mots apparaissant à gauche et à droite du point de focus. Un minimum local de cette mesure est considéré comme l'indice d'une zone de changement thématique, et une frontière est alors définie à l'emplacement de la limite de paragraphe la plus proche. Les mots ayant joué un rôle important dans le maintien à une valeur élevée de la mesure de cohérence lexicale entre deux minima sont mis à profit pour caractériser le thème de la région considérée. Si l'algorithme décrit permet en effet une segmentation thématique pertinente à l'échelle de groupes de paragraphes, rien ne garantit que deux zones non contiguës traitant d'un même thème soient caractérisées par des listes de mots identiques, ce qui rend difficile la détection de proximité thématique entre segments non consécutifs.

Le couple de traitements SEGCOHLEX / SEGAPSITH présenté dans [FG01] est pour sa part basé sur une idée similaire mais réalise une segmentation de granularité beaucoup plus fine, la mesure de cohésion lexicale étant calculée pour chaque occurrence de mot et à l'aide de fenêtres de 10 mots à gauche et à droite du point de focus. Cet indice faisant usage d'une quantité de mots bien plus faible que le précédent, des données supplémentaires sont utilisées pour accroître la quantité d'information qu'il cristallise : un corpus de 45 millions de mots est employé au préalable afin d'extraire un réseau de collocations, et cette connaissance est exploitée dans la mesure de proximité lexicale pour réaliser une première segmentation des textes étudiés (module SEGCOHLEX). À l'aide de cette segmentation, le système définit des « signatures thématiques » qui forment la base d'une seconde segmentation (module SEGAPSITH) et fournissent une caractérisation indirecte mais, contrairement à TEXTTILING, systématique des thèmes détectés.

Notre recherche se distingue de ce dernier travail de par la définition même de la problématique abordée : le système FAESTOS présenté dans ce chapitre a pour but de s'intégrer de manière transparente à une chaîne de traitements dont l'objectif est l'acquisition automatique de lexiques sémantiques. Il doit donc fonctionner de façon totalement autonome, sans requérir l'apport de données complémentaires. De plus, nous préservons le découpage en paragraphes du corpus initial ; nous ne nous intéressons pas à une structuration de la lecture séquentielle des textes, mais cherchons à construire une caractérisation permettant de connaître de manière directe le ou les thèmes abordé dans un segment textuel défini *a priori*. Nous n'effectuons donc pas un parcours analytique linéaire du corpus, mais traitons « en bloc » un ensemble de données statistiques collectées sur celui-ci. Enfin, pour satisfaire notre double objectif de détection et caractérisation de thèmes, il nous faut fournir une représentation de ceux-ci aisément appréhendable par un utilisateur humain afin que les informations lexicales extraites soient exploitables.

Le travail présenté par G. Salton *et al.* dans [SSBM96] aborde une problématique plus directement apparentée à la nôtre, mais fonctionne à une échelle plus restreinte : son objectif consiste à rassembler les paragraphes d'un article journalistique ou scientifique en fonction des thèmes qu'ils abordent. Les auteurs réalisent dans un premier temps un regroupement de paragraphes successifs selon une méthode similaire à celle employée par TEXTTILING, puis rapprochent les groupes non contigus par un critère de similarité lexicale calculé à l'échelle de paragraphes entiers. Ces informations de structures sont utilisées afin de guider la sélection des passages d'un article les plus pertinents en réponse à une requête documentaire. La problématique de caractérisation des thèmes sous-tendant le regroupement des paragraphes n'est pas abordée, ce qui limite l'intérêt direct de la méthode employée dans le cas qui nous préoccupe ; de plus, le système présenté est destiné à l'étude d'articles individuels et n'est pas conçu pour supporter le passage à l'échelle d'un corpus de plusieurs millions de mots. Nous faisons néanmoins usage au cours des traitements

développés ici d'une mesure de similarité lexicale entre paragraphes apparentée à celle employée par G. Salton *et al.*

De nombreux travaux, enfin, obtiennent de très bons résultats grâce à l'emploi d'outils plus élaborés pour la détection de thèmes : dans [Bru03], A. Brun fait ainsi usage de plusieurs modèles statistiques de langage (grammaires n-grammes et modélisation par « mémoire cache » [KdM90]) afin de caractériser le type de langue associé à chaque thème. L'apprentissage de ces modèles étant réalisé de manière supervisée sur des ensembles de textes de thèmes connus, il est plus difficile dans ce cas de faire une comparaison avec nos travaux. On peut néanmoins noter que, de manière générale, l'inconvénient d'une approche supervisée est de fixer *a priori* les thèmes recherchés alors que nous les faisons émerger des données textuelles elles-mêmes, ce qui permet une meilleure prise en compte des spécificités du corpus étudié.

Avant de décrire de manière détaillée le système FAESTOS, nous procédons à la section suivante à quelques mises au point pratiques et théoriques concernant le contexte de notre travail.

2.3. Précisions introductives

Cette section a pour but de fournir quelques éléments d'information nécessaires à la compréhension de ce chapitre. Nous y présentons tout d'abord plus en détail le corpus d'étude choisi, ainsi que l'analyse préalable à laquelle il a été soumis. Nous introduisons ensuite les concepts de la classification ascendante hiérarchique (CAH), outil d'analyse statistique de données que nous avons largement employé au cours de nos travaux. Enfin, nous décrivons rapidement les principes de CHAVL (Classification Hiérarchique Ascendante par Analyse de la Vraisemblance du Lien), méthode de CAH à laquelle nous faisons plus particulièrement appel dans ce chapitre.

2.3.1. Corpus d'étude

Le corpus employé au cours de notre recherche afin d'expérimenter les différentes méthodes développées et d'évaluer leur efficacité est composé de quatorze ans d'archives du mensuel *Le Monde diplomatique*, échelonnées de 1985 à 1998. *Le Monde diplomatique* est une publication orientée vers la réflexion, qui fait le choix d'une prise de distance souvent idéologiquement marquée, mais toujours amplement argumentée, par rapport à l'actualité immédiate. Comparativement à un journal d'information classique, les articles sont en moyenne plus longs et développent un discours moins factuel et plus « sophistiqué », dans une langue recherchée qui ne rechigne pas à la complexité — par rapport à certains corpus usuellement employés pour des travaux d'acquisition automatique d'informations lexicales, *Le Monde Diplomatique* constitue sans conteste un réel défi…

Les thématiques abordées dans ce corpus relèvent pour la plupart de considérations politiques, géopolitiques ou macroéconomiques, mais laissent également une bonne place à la culture (évocation d'œuvres particulières ou discussions d'ordre plus général sur l'évolution des courants artistiques) et à l'analyse sociologique.

Après cette introduction qualitative, nous présentons dans ce qui suit la composition du corpus d'un point de vue plus technique : dans un premier temps, son volume et la répartition de ses unités textuelles, puis son mode de préparation à l'analyse numérique.

2.3.1.1. Composition du corpus

Notre corpus rassemble 5 704 articles, choisis suivant leur type afin de rassembler l'essentiel du contenu réellement journalistique des archives disponibles, par opposition notamment au courrier des lecteurs ou à l'annonce du sommaire. Les catégories retenues sont :

- les articles « classiques » (4 199 exemplaires) : il s'agit le plus souvent d'articles de fond d'une à deux pages, qui rassemblent en moyenne 2 000 à 2 500 mots en 20 paragraphes ;
- les encadrés (1 013 exemplaires) : ce sont de courts textes explicatifs accompagnant un article ; ils rassemblent en moyenne 400 mots dans 4 à 5 paragraphes ;
- les reportages (376 exemplaires) : plus factuels et plus longs que les articles simples, ils rassemblent en moyenne 3 200 mots dans 26 paragraphes ;
- les éditoriaux (116 exemplaires) : textes de longueur intermédiaire introduisant chaque numéro du mensuel, en moyenne quelque 1 400 mots regroupés en 13 paragraphes.

Le corpus étudié se présente donc comme une collection de textes riches, plus longs et approfondis en moyenne que ceux, par exemple, d'un quotidien d'information. Dans sa totalité, le corpus rassemble 98 432 paragraphes[1], soit 11 380 197 occurrences de lexies (les termes complexes les plus courants, noms propres en plusieurs parties, *etc.*, étant considérés comme une seule unité). La figure 2.1 illustre la répartition des tailles de paragraphes dans le corpus.

2.3.1.2. Prétraitement du corpus

Afin de permettre son exploitation, le corpus a été segmenté, étiqueté catégoriellement et lemmatisé ; il n'a en revanche pas subi d'analyse syntaxique, même superficielle.

[1] Seuls les paragraphes de plus de vingt mots ont été retenus, ce qui exclut notamment les titres d'article, sous-titres, et signatures ; les notes de bas de page ont également été retirées, ainsi que les tableaux de données.

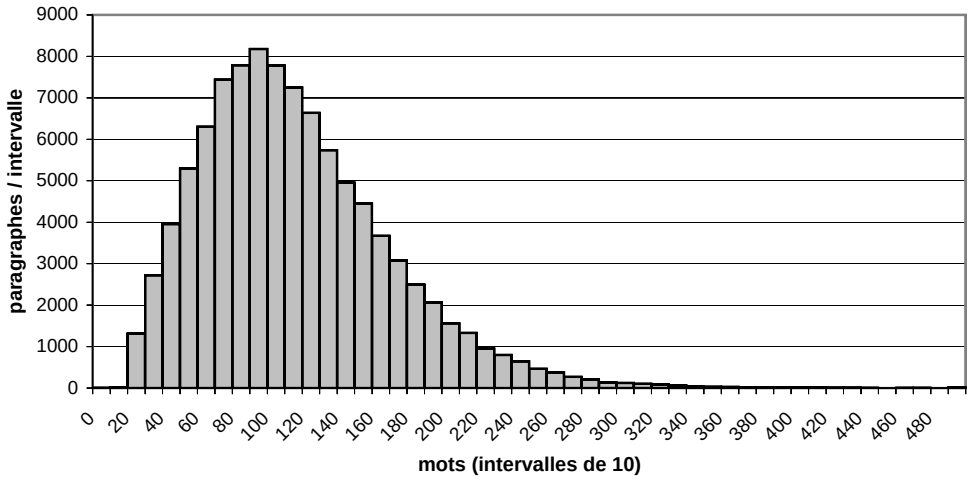


FIG. 2.1 – *Répartition des tailles de paragraphes dans le corpus d'archives du* Monde diplomatique

Segmentation. Il s'agit du découpage du flux textuel en entités lexicales[2]. Elle a été réalisée grâce à l'outil mtseg du projet Multext [Arm96]. Celui-ci fait usage de règles définies sous la forme d'une grammaire rationnelle, ainsi que de listes d'exceptions. Nous avons été amené par la suite à modifier cette segmentation grâce à des scripts Perl *ad hoc*, permettant par exemple le repérage et l'agglomération des parties de noms propres composés ou de dates.

Étiquetage et lemmatisation. L'étiquetage consiste à associer à chaque occurrence de mot du corpus des informations concernant sa catégorie morpho-syntaxique (nom, verbe, adjectif, *etc.*) et son état flexionnel (genre, nombre, temps et mode des verbes, *etc.*). Par la lemmatisation, on ramène l'occurrence à sa forme canonique conventionnelle (infinitif pour les verbes, masculin singulier pour les adjectifs, *etc.*). Ces opérations ont été effectuées conjointement, en deux étapes. Dans un premier temps, l'outil d'étiquetage mmorph du projet Multext a été employé afin d'associer à chaque *token* l'ensemble de ses étiquettes possibles, ainsi que les lemmes associés. Contrairement à de nombreux autres outils d'étiquetage catégoriel [Bri92], mmorph ne « devine » pas la catégorie d'un mot par analyse morphologique, mais fait usage de listes de mots connus étiquetés à partir desquelles il engendre toutes les formes possibles grâce à un moteur de génération morphologique et un ensemble de règles linguistiques. L'opération d'étiquetage consiste ensuite à trouver pour chaque mot toutes ses occurrences dans la base de données morphologique ainsi construite. Cette approche présente l'avantage de la fiabilité, son prin-

[2]La tâche de segmentation peut également comprendre une détection des limites de phrase, que nous ne mentionnons pas car l'outil utilisé se contente de marquer comme telles toutes les ponctuations fortes rencontrées.

cipal défaut étant de requérir un important travail d'adaptation du lexique original au corpus traité : nous avons ainsi étendu le lexique d'environ 30 000 mots livré avec le programme pour atteindre quelque 60 000 mots, dont beaucoup sont très spécifiques au corpus étudié.

Les nombreuses ambiguïtés d'étiquetage apparues ont dans un deuxième temps été levées grâce au logiciel Tatoo [ABR95], développé à l'ISSCO de Genève, qui réalise cette tâche par l'utilisation d'un modèle de Markov caché dont les états sont les catégories morphosyntaxiques et les symboles les classes d'ambiguïtés observées dans le corpus. Les probabilités de transitions entre états (catégories morphosyntaxiques) ainsi que les probabilités d'association entre ceux-ci et les symboles (classes d'ambiguïtés) peuvent être évaluées de manière non supervisée par l'algorithme de Baum-Welch ou apprises plus précisément en exploitant une portion de texte désambiguïsée manuellement. Il est également possible de régler plus finement le fonctionnement de l'algorithme par l'intermédiaire de biais spécifiés manuellement.

Nous précisons maintenant la forme sous laquelle nous considérons les données textuelles pour notre travail d'extraction automatique de classes de mots-clés.

2.3.1.3. Préparation du corpus à l'analyse numérique

Comme nous l'avons annoncé en introduction, le paragraphe est considéré dans le cadre du système présenté dans ce chapitre comme une unité textuelle « atomique », à l'échelle de laquelle se développent les phénomènes d'isotopie que nous souhaitons mettre au jour. Selon ce choix, la structure interne des paragraphes doit nous rester « invisible » ; nous les considérons donc comme des « sacs » de mots, l'unique donnée exploitable étant le nombre d'occurrences de chaque lemme dans chaque paragraphe du corpus, autrement dit la répartition des mots du corpus sur ses paragraphes. Pour nos besoins, nous assimilons donc le corpus à une matrice $\mathcal{M}$ dont chaque ligne correspond à un lemme de mot, et chaque colonne à un paragraphe. La valeur présente dans la matrice à la ligne i, colonne j est le nombre d'occurrences du lemme i sous l'une quelconque de ses formes dans le paragraphe j. Les mots du corpus apparaissant en moyenne 100 fois environ dans celui-ci, un rapide calcul montre que plus de 99,9 % des valeurs de la matrice sont nulles : $\mathcal{M}$ est très « creuse ». Cette propriété est très handicapante pour la plupart des méthodes statistiques d'analyse de données[3], et nous amènera à développer des techniques destinées en particulier à la résolution de ce problème.

Cette forme de représentation matricielle des données a pour intérêt de favoriser grandement leur analyse statistique, en particulier par des méthodes

[3]Certains travaux s'intéressent d'ailleurs en particulier à l'adaptation de méthodes de classification « classiques » au cas de matrices très creuses [JNXH05], mais ils se situent majoritairement dans le domaine de la catégorisation de documents, ce qui les rend difficilement applicables dans le cadre de notre problématique.

de classification ascendante hiérarchique dont nous faisons largement usage dans le système FAESTOS, et dont nous présentons maintenant le principe.

2.3.2. Classification ascendante hiérarchique

L'objectif de la classification non supervisée est le regroupement d'un ensemble d'objets en classes construites de manière à maximiser leur cohérence pour une certaine mesure de similarité. Comme il n'est pas envisageable, pour une collection d'objets importante, de procéder à une recherche exhaustive sur toutes les partitions qu'il est possible de définir sur cette collection afin de trouver un optimum absolu de cohérence des classes, de nombreux algorithmes ont été développés afin d'aboutir à une partition quasi optimale pour un coût calculatoire modéré. La classification ascendante hiérarchique (CAH) est l'une de ces méthodes, et sans doute l'une des plus largement utilisées. Elle présente notamment l'avantage de ne pas requérir de présupposé concernant le nombre ou la taille des classes d'objets à produire, ce qui est capital pour notre travail, mais également d'être d'un principe suffisamment simple pour qu'il soit aisé d'interférer au besoin avec son fonctionnement.

Étant donné un ensemble d'objets à classer $\mathcal{O}$ et une mesure de similarité s reflétant de manière numérique la ressemblance entre objets de $\mathcal{O}$ (la plupart des méthodes de classification exigent que s soit symétrique, et considèrent en conséquence ses valeurs rassemblées dans une demi-matrice), l'objectif de la CAH est de construire un arbre de classification du type de celui de la figure 2.4, page 58. Les feuilles de l'arbre correspondent aux objets à classer, et chaque nœud matérialise une classe résultant de la fusion entre les classes associées à ses fils. Ainsi le nœud marqué *(b)* sur la figure 2.4 correspond-il à la classe {*cinéma*, *film*, *scène*, *ville*}.

La construction de l'arbre se déroule comme suit : partant d'une partition initiale triviale de $\mathcal{O}$, où chaque objet de l'ensemble est placé dans un singleton (les feuilles de l'arbre), on construit une famille de partitions de plus en plus « généralisantes », chacune d'entre elles étant obtenue à partir de la précédente en fusionnant les deux classes de cette dernière présentant une similarité maximale (cette fusion est matérialisée par un nœud de l'arbre). Le processus s'arrête lorsque la totalité des objets est rassemblée en une unique classe (la racine de l'arbre). Ainsi l'arbre de classification reflète-t-il, à mesure que l'on remonte des feuilles à la racine, le déroulement temporel de ce processus d'agrégations successives, chacun de ses niveaux correspondant à une étape de l'algorithme.

La simplicité apparente du procédé décrit cache une difficulté réelle : la nécessité d'exprimer la proximité entre deux classes en fonction de la similarité s définie entre objets. Il est par exemple possible de considérer pour cela que la similarité entre deux classes est égale à la moyenne des similarités entre leurs éléments (« lien moyen »), au maximum de ces similarités (« lien maximal »), ou à leur minimum (« lien complet »). D'autres approches s'appliquent uni-

quement à certains types de données : le classique « indice de Ward » [War63], par exemple, minimise la perte d'inertie induite dans le système constitué de l'ensemble des points représentant les objets à classer dans un repère euclidien lorsqu'on remplace ces points par les centres des classes formées ; permettant d'obtenir des résultats d'assez bonne qualité, il n'est néanmoins adapté qu'au cas d'objets caractérisés par des attributs numériques vectoriels, qu'il est possible de représenter comme un nuage de points.

Nous avons choisi pour nos travaux de faire appel à la méthode CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien) développée par I. C. Lerman [Ler70, Ler91, LLP94]. Celle-ci présente deux avantages importants pour nous : d'une part, elle propose un mécanisme de calcul de similarités entre classes totalement indépendant du type de données étudié, et adaptable à tous types de mesures de similarité. De plus, cette méthode est inspirée de principes probabilistes issus de la théorie de l'information, ce qui semble particulièrement adapté au champ de notre recherche étant donné le succès qu'ont pu rencontrer en TAL des outils et mesures comme l'information mutuelle*, l'entropie* ou la divergence de Kullback-Leibler*, issus de cette même théorie. La section suivante est consacrée à l'exposé rapide de ses principes.

2.3.3. CHAVL : une méthode de CAH

CHAVL adopte pour résoudre le problème du passage d'une mesure de similarité entre objets à une similarité entre classes une vision probabiliste du concept de similarité, en convertissant l'indice de similarité brut initial en une probabilité d'existence de lien entre objets (d'où la dénomination « analyse de la vraisemblance du lien » — AVL). Nous présentons dans ce qui suit la méthodologie mathématique mise en œuvre afin d'opérer ce changement de point de vue.

La vraisemblance de l'existence d'un lien entre deux objets est d'autant plus grande que leur similarité est « invraisemblablement élevée » au regard des autres paires d'objets. Elle est évaluée selon le principe suivant : $\mathcal{O}$ désigne toujours l'ensemble des objets à classer, s la mesure de similarité employée et $a, b \in \mathcal{O}$ deux objets à classer ; α et β étant deux objets quelconques dans $\mathcal{O}$, on calcule la probabilité que $s(\alpha, \beta)$ ait une valeur supérieure à $s(a, b)$. Formellement, cette probabilité s'exprime par

$$p(a, b) = Pr_{\alpha,\beta \in \mathcal{O}} \left(s(\alpha, \beta) > s(a, b) \right)$$

et se calcule en faisant l'hypothèse que s obéit à une loi normale, l'écart type et la moyenne de s étant évalués sur ses valeurs constatées entre toutes les paires d'objets à classer. Sous une hypothèse d'absence de lien (h.a.l.) entre a et b, on devrait observer $p(a, b) = 0,5$. Plus $p(a, b)$ est faible devant cette valeur, plus l'h.a.l. est improbable, c'est-à-dire plus l'existence d'un lien entre a et b

est vraisemblable. On substitue donc à la demi-matrice initiale des similarités entre objets la demi-matrice des $P(a,b)$ telles que

$$P(a,b) = 1 - p(a,b) = Pr_{\alpha,\beta \in \mathcal{O}} \left(s(\alpha,\beta) \leq s(a,b) \right)$$

La proximité entre deux classes d'objets se calcule selon un principe similaire : étant données deux parties C et D de $\mathcal{O}$ disjointes, on considère la valeur du lien maximal entre C et D, défini par :

$$q(C,D) = \max_{(c,d) \in C \times D} \left(P(c,d) \right)$$

qui est égal au maximum des vraisemblances de liens entre éléments de C et D. De même que ci-dessus, on pose une h.a.l. entre C et D et l'on évalue la probabilité $Q(C,D)$ qu'il existe sous cette hypothèse d'avoir, pour Γ et Δ deux parties quelconques de $\mathcal{O}$ de cardinaux respectivement égaux à ceux de C et D, $q(\Gamma,\Delta) \leq q(C,D)$. Une fois encore, plus cette probabilité $Q(C,D)$ approche 1, plus l'existence d'un lien entre C et D est vraisemblable. Cette valeur, dont il est montré dans [Ler91] qu'elle se calcule par :

$$Q(C,D) = q\left(C,D\right)^{lm}$$

où l et m sont les cardinaux respectifs de C et D, est appelée « vraisemblance du lien maximal ». La formule réellement employée par CHAVL pour calculer le lien existant entre C et D est une variante de cette mesure, définie par :

$$VL_{\varepsilon}(C,D) = q\left(C,D\right)^{(lm)^{\varepsilon}}$$

Le paramètre additionnel ε permet de faire varier de manière progressive la mesure depuis un critère simple de maximum des liens entre objets des classes considérées ($VL_0(C,D) = q(C,D)$) jusqu'à la formule de vraisemblance du lien maximal ($VL_1(C,D) = Q(C,D)$). Ce paramètre permet donc de contrôler dans une certaine mesure la stratégie de regroupement de classes adoptée par CHAVL, et d'adapter ainsi le fonctionnement de l'algorithme aux données traitées.

Utilisation de CHAVL

Les principes de la méthode CHAVL sont implantés dans le programme `chavleps`, implémenté en FORTRAN. Ce programme propose en outre un ensemble de mesures de similarité prédéfinies fondées sur l'AVL et adaptées aux types de données les plus couramment rencontrés : matrice d'incidence, vecteurs d'attributs quantitatifs, tableau de contingence, *etc.* `chavleps` permet également de spécifier à l'exécution la valeur du paramètre ε employé dans la formule de calcul du lien entre classes.

L'une des difficultés résultant de l'emploi d'une méthode de classification hiérarchique telle que CHAVL est la nécessité de passer d'une représentation

structurée sous forme d'arbre à une partition en classes de l'ensemble des objets étudiés, dont l'exploitation informatique est plus immédiate. La méthode la plus couramment employée pour ce faire consiste à choisir un niveau dans l'arbre et à « couper » l'ensemble de ses branches à cette hauteur ; on retrouve ainsi l'une des partitions générées au cours de l'algorithme d'agrégations successives. La ligne pointillée de la figure 2.4 désigne un tel niveau de découpage, qui génère la partition {{*cinéma, film, scène, ville, auteur, œuvre, logement, cité*}, {*capitalisme, vie*}, {*chômage, emploi, chômeur*}}.

Ce niveau peut être choisi en fonction de contraintes imposées sur le nombre de classes souhaitées ou leur taille moyenne, mais cela équivaut à renoncer à l'intérêt principal de la classification hiérarchique, qui est justement de ne pas requérir de présupposé de cette nature. CHAVL propose donc un indicateur statistique associant à chaque niveau de l'arbre une évaluation de la cohérence moyenne des classes formées à cette hauteur, un maximum local de cette mesure correspondant à une partition particulièrement intéressante — on parle de « niveau significatif » de l'arbre. Cette indication n'est néanmoins pas toujours suffisante, et il peut être nécessaire, comme dans notre cas, de mettre en œuvre une stratégie plus « intelligente » d'extraction des classes les plus pertinentes de l'arbre de classification.

Ayant présenté ces éléments de contexte de notre recherche, nous pouvons désormais procéder à la description du système FAESTOS. Nous introduisons tout d'abord le principe de la caractérisation et détection de thèmes par classes de mots-clés, en partant du fondement linguistique de ce concept en sémantique différentielle, avant d'exposer la structure générale du système que nous avons mis en œuvre afin de réaliser leur apprentissage de manière totalement automatique.

2.4. Principes de la méthode de caractérisation et détection de thèmes développée

Le fonctionnement du système FAESTOS se fonde sur l'utilisation des *isotopies sémantiques* de F. Rastier, présentées au chapitre précédent. Visant l'échelle sémantique du thème, que nous mettons en correspondance avec le niveau du domaine dans la structuration lexicale, il est naturellement pertinent de s'attacher à la détection des isotopies mésogénériques ; nous avons pu montrer (figure 1.3, page 34) que le paragraphe était bien un niveau de granularité dans le découpage du texte pertinent pour l'observation de ces isotopies.

L'implantation informatique de ce principe semble *a priori* problématique : en effet, ne disposant d'aucune information sémantique sur le corpus ou les mots qui le composent, nous nous trouvons dans l'impossibilité de faire appel aux sèmes portés par les mots apparaissant dans un paragraphe pour tenter d'y détecter une isotopie. Une idée permettant de surmonter cet obstacle est

proposée par R. Pichon et P. Sébillot dans [PS00], travail où le système présenté ici trouve ses racines. Son principe consiste à approcher la notion d'isotopie sémantique par celle de coprésence de mots particulièrement symptomatiques d'un thème, appelés *mots-clés*. Le texte de la figure 1.3 fournit une illustration de cette notion : si l'on en extrait des termes comme *réseau*, *ordinateur*, *Internet* ou *télécommunication*, il n'est en effet nul besoin d'en savoir davantage sur leur contexte d'apparition pour avoir une forte présomption concernant le thème (pas forcément unique) développé. On peut en déduire que ces termes, apparaissant de manière conjointe, exercent les uns sur les autres des contraintes interprétatives suffisamment fortes pour que l'on puisse être assuré de voir s'exprimer les sèmes prenant part à l'isotopie, quel que soit le contexte environnant[4]. Si l'on définit un mot-clé comme un terme de forte « saillance thématique » associé à un thème donné, la présence conjuguée de plusieurs de ces mots-clés induit donc une probabilité élevée de voir apparaître l'isotopie sémantique révélatrice de ce thème.

Notons encore que la présence d'un unique mot-clé dans un segment de texte donné (un paragraphe dans le cadre de nos travaux) ne fournit aucune indication fiable : en effet, à défaut de connaître les sens des mots apparaissant conjointement avec lui dans ce paragraphe, il est impossible d'anticiper la manière dont il devra y être interprété[5]. Inversement, plus les mots-clés sont nombreux, plus le diagnostic de présence du thème est fiable, non seulement, comme on peut le comprendre intuitivement, par accumulation d'indices convergents, mais aussi grâce à la sûreté d'interprétation que fournissent les nombreuses contraintes sémantiques exercées par ces mots les uns sur les autres. Nous adoptons néanmoins dans un premier temps un critère minimal de ce point de vue, puisque nous n'exigeons que la présence conjuguée de deux mots-clés pour affirmer l'apparition d'un thème.

Dans [Ras96], F. Rastier estime qu'environ 5 à 10 % des mots d'un texte participent à une isotopie ; dans le cas des paragraphes de notre corpus (qui contiennent une centaine de mots en moyenne), on attendrait donc de 5 à 10 mots participant à l'isotopie dans chaque paragraphe. Pourquoi alors se contenter de deux mots-clés ? La figure 2.2 illustre pour répondre à cette question la différence entre la notion théorique d'isotopie et celle appliquée de mots-clés : la première voit le jour par la répétition de sèmes, potentiellement portés par des sémèmes en eux-mêmes relativement anodins (*ligne*) ou susceptibles de multiples interprétations (*réseau*). Les mots-clés, en revanche, s'ils participent tous à l'isotopie, ne doivent laisser de possibilités d'ambiguïtés que minimales, et il n'est donc pas pertinent de considérer comme mots-clés des mots pour lesquels les sèmes participant à l'isotopie ne sont qu'afférents et actualisés de manière très spécifique par le contexte d'usage. Enfin, d'un point

[4] Il est bien entendu possible de trouver un contre-exemple infirmant cette affirmation, mais nous sommes contraint par souci d'efficacité de faire abstraction des cas « pathologiques ».

[5] Il existe des mots ne présentant aucune ambiguïté sémantique, mais nous ne disposons pas d'informations nous permettant d'affirmer *a priori* qu'un mot donné entre dans cette catégorie.

« En fait, les réseaux électroniques sont majoritairement le domaine réservé des élites et de leurs satellites - principalement les classes moyennes des pays du Sud, en Corée, en Inde ou au Brésil. Pour les privilégiés, **naviguer** dans les réseaux est le signe d'un **progrès** capital. Mais considérons les faits : en 1995, le nombre d'ordinateurs personnels (*PC*) en usage dans le monde était d'environ 180 millions , pour une population globale de presque six milliards d'individus. La possibilité d'**accéder** au réseau était donc limitée à 3 % de personnes. En 1995, un petit nombre de pays riches, représentant environ 15 % de la population mondiale, possédaient environ les trois quarts des principales **lignes téléphoniques**, sans lesquelles, par le biais d'un *modem* (*modulateur-démodulateur*) on ne peut accéder à Internet... Plus de la moitié de la planète ne s'était jamais servie d'un **téléphone** : dans quarante-sept pays, il n'y avait même pas une seule **ligne** pour cent habitants ; or, d'après l'Union internationale des télécommunications, on peut affirmer qu'il dessert les besoins d'un pays lorsque nul n'est situé à plus de cinq kilomètres d'un **téléphone**. »

FIG. 2.2 – *Paragraphe identique à celui de la figure 1.3, sur lequel nous avons fait apparaître les mots participant à l'isotopie domaniale « /communication/ + /distance/ + /électronique/ ». Les mots sur fond gris participent à l'isotopie et peuvent faire usage de mots-clés ; néanmoins, les mots en italique sont trop rares pour être extraits par une analyse statistique. Les mots en gras, pour leur part, participent à l'isotopie mais sont de sens trop général pour être retenus comme mots-clés du domaine*

de vue plus pratique, les mots-clés doivent être suffisamment fréquents pour donner prise aux méthodes d'apprentissage par analyse statistique de données que nous mettons en œuvre. La notion de classe de mots-clés est donc bien plus restrictive que celle de domaine, et s'ils appartiennent bien au domaine considéré, les mots mis en relief dans l'exemple de la figure 2.2 ne sont pas tous des mots-clés. On constate qu'à l'évidence le problème à traiter est autant une question de *sélection* de mots pertinents pour la détection et caractérisation thématique que de construction de classes thématiques. L'architecture du système mis en œuvre, que nous présentons maintenant, est conçue afin de répondre à cette double problématique.

L'ensemble de traitements présenté a pour objectif de générer de manière automatique à partir des seules données textuelles rassemblées dans un corpus des classes de mots-clés reflétant ses principaux thèmes. Ces classes doivent répondre à un double objectif : d'une part, permettre de détecter de manière efficace les apparitions de « leur » thème dans le corpus et, d'autre part, constituer une description aisément lisible et compréhensible par un humain du thème qu'elles caractérisent. L'idée retenue pour construire ces classes consiste à considérer que, puisque les mots que l'on cherche à rassembler sont révélateurs d'un même thème dont leur présence conjuguée est un indicateur fort, il existe une forte probabilité de les voir apparaître dans les mêmes paragraphes. Partant de cette idée, le système FAESTOS réalise en trois étapes successives la construction des classes de mots-clés, en faisant usage de la seule donnée de répartition des mots du corpus sur ses paragraphes. Afin de faciliter la compréhension de la description du système, souvent assez technique, nous présentons ici dans un premier temps le principe général de ses trois étapes, qui sont détaillées dans les sections suivantes. La structure générale de FAESTOS est présentée de manière synthétique sur la figure 2.3, qui peut servir de référence et de guide de lecture pour la suite.

Nous effectuons dans un premier temps une classification des mots grâce à la méthode de classification ascendante hiérarchique CHAVL, afin de rassembler les mots présentant des répartitions similaires sur les paragraphes du corpus d'étude. Ses résultats sont exploités grâce à un algorithme destiné à extraire de manière « intelligente » des classes de l'arbre de classification produit en prenant en compte les spécificités de la problématique traitée. Cette étape, détaillée en section 2.5, permet l'organisation en classes d'une sélection de mots du corpus choisis selon leur type grammatical (noms et adjectifs) et leur fréquence[6].

Ce mode de sélection est toutefois peu satisfaisant : comme nous l'avons vu, les noms et adjectifs du corpus n'ont pas tous vocation à jouer le rôle de mots-clés, et les classes obtenues sont donc « parasitées » par de nombreux

[6]L'analyse statistique des répartitions des mots ne peut en effet être significative qu'au-delà d'un certain nombre d'occurrences.

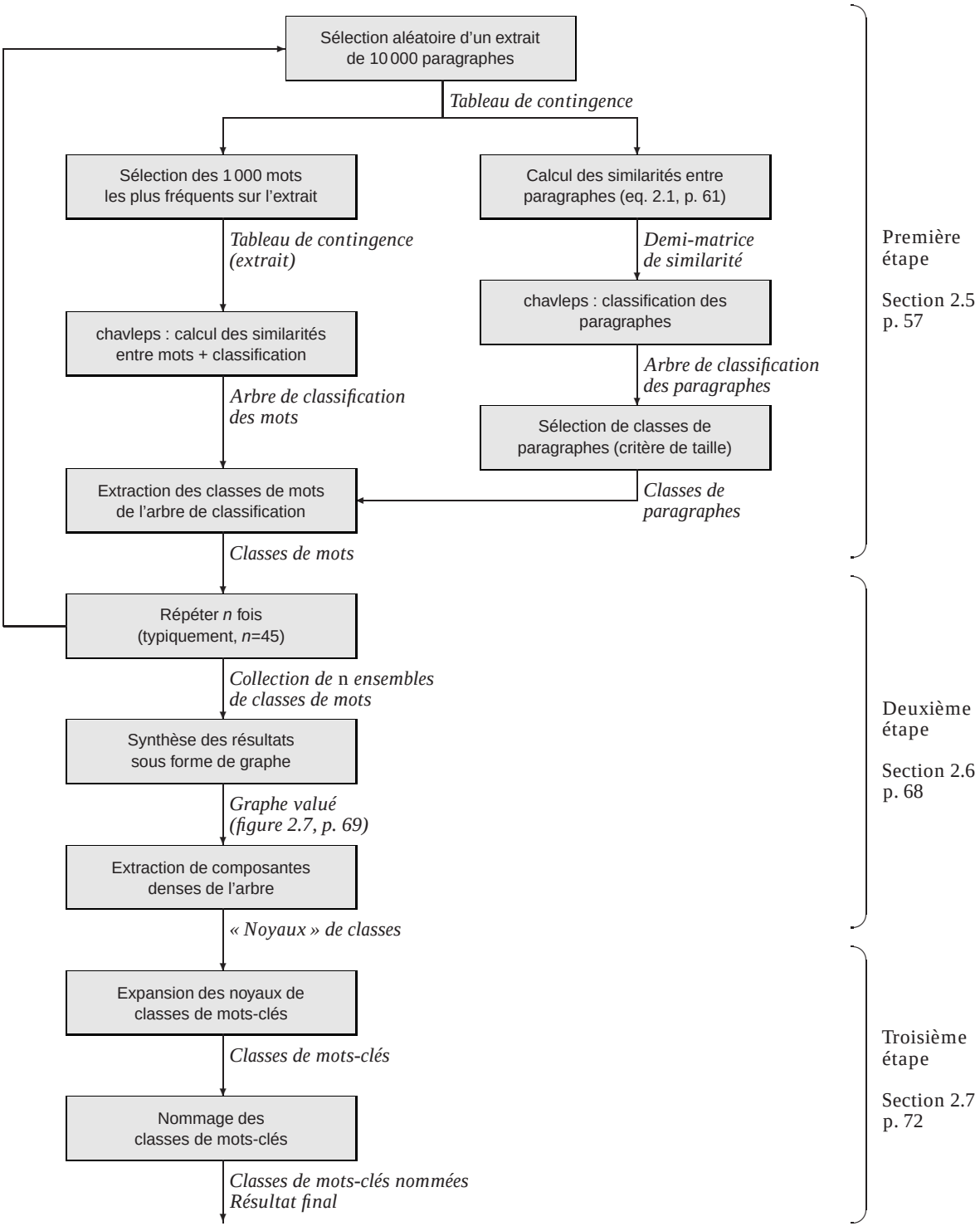


FIG. 2.3 – *Organisation des opérations du programme d'extraction de classes de mots-clés*

mots non pertinents pour la détection de thèmes. Inversement, de nombreux mots fortement révélateurs d'une thématique sont trop peu fréquents pour donner prise à une analyse statistique. Les étapes suivantes ont pour objectif de pallier ces insuffisances, incontournables dans un premier temps. Elles permettent en outre de surmonter une difficulté plus technique : la complexité des calculs effectués nous contraint en effet à ne travailler que sur un extrait du corpus sélectionné aléatoirement (environ 10 % du texte)[7] ; or on constate, au vu des résultats obtenus, qu'il est nécessaire de prendre en compte l'intégralité des données textuelles disponibles. En effet, l'extrait étudié n'est pas forcément représentatif de l'ensemble du corpus, et les thèmes présents dans celui-ci ne sont donc pas toujours tous suffisamment représentés dans l'extrait pour susciter l'émergence d'une classe de mots-clés. De plus, afin que la détection de thèmes réalisée grâce aux classes de mots-clés produites ne présente aucune ambiguïté, il est impossible de permettre le partage de mots par plusieurs classes. Les mots pouvant se trouver associés à plusieurs thèmes, et notamment les mots polysémiques, doivent donc pouvoir être placés dans la classe de mots-clés correspondant à leur sens le plus courant dans l'ensemble du corpus.

La seconde étape mise en œuvre, présentée en section 2.6, permet une « remontée » de l'analyse à l'échelle du corpus entier. Elle considère pour cela les résultats de plusieurs études partielles réalisées sur des extraits aléatoires distincts et les combine en associant les mots selon leurs regroupements les plus fréquents. La technique employée, qui ne retient les associations entre mots que si elles sont observées dans plusieurs des classifications réalisées sur des extraits du corpus, permet également d'écarter de la classification finale les mots très polysémiques ou peu marqués thématiquement, qui ne participent à aucun regroupement stable. Enfin, en requérant le consensus de plusieurs classifications partielles pour rassembler deux mots, elle réalise un filtrage du « bruit » existant dans les résultats de la première étape de calcul, bruit par définition non répétitif. On obtient ainsi un nouvel ensemble de « noyaux » de classes contenant chacun quelques mots très fortement associés à un thème particulier.

Ces classes permettent d'effectuer une première détection de thèmes sur l'ensemble du corpus, et de réaliser à partir de ce résultat un apprentissage de type supervisé prenant en compte les mots trop peu fréquents pour avoir pu être considérés auparavant. Cette dernière étape de construction, décrite à la section 2.7, engendre des classes d'une trentaine de mots en moyenne rendant assez fidèlement compte de l'étendue du « champ lexical » associé à chacun des principaux thèmes du corpus étudié. Une heuristique simple permet enfin d'extraire de chacune de ces classes un triplet de mots particulièrement représentatifs, utilisé pour la « nommer » et faciliter son exploitation ultérieure.

[7] Les difficultés rencontrées sont, d'une part, le temps de calcul, d'autre part, les aberrations pouvant naître lors de calculs en « virgule flottante » du fait de la très faible densité des données étudiées.

La section 2.8 présente les résultats obtenus grâce au système FAESTOS, dont nous proposons plusieurs évaluations suivant différents critères. Nous évoquons en section 2.8.4 quelques évolutions envisageables pour améliorer ses performances, notamment les possibilités de complémentarité avec des travaux envisageant sous d'autres points de vue le même problème de détection de thèmes, puis revenons à notre préoccupation initiale de construction de lexiques : la section 2.9 opère un retour sur l'ancrage linguistique de notre travail et présente la manière dont, ayant caractérisé et détecté les thèmes du corpus, nous découpons celui-ci en sous-corpus thématiques et extrayons le vocabulaire spécifique à chacun d'eux, qui forme le domaine associé.

2.5. Première étape : analyse statistique des répartitions

Le principe fondamental du premier traitement mis en œuvre afin de parvenir à la construction automatique de classes de mots-clés représentatifs des principaux thèmes abordés dans le corpus est inspiré du travail antérieur de R. Pichon et P. Sébillot [PS00] déjà mentionné. Afin de fournir un point de comparaison, nous décrivons ici dans un premier temps la méthodologie et les résultats de cette recherche, pour nous concentrer dans la suite de cette section sur les évolutions que nous avons fait subir à cette base.

2.5.1. Héritage du travail précurseur

Les auteurs de [PS00] travaillent comme nous à l'échelle de paragraphes de textes entiers, et formulent également l'hypothèse que les proximités thématiques entre mots peuvent se déduire des similarités existant dans leurs répartitions parmi les paragraphes du corpus. Ils se restreignent pour la détection de thèmes à l'étude des noms, couramment considérés comme porteurs privilégiés du contenu informationnel, et donc thématique, d'un texte [Sme99]. Exploitant pour leur étude le même corpus du *Monde Diplomatique* que nous avons retenu, ils choisissent afin de contourner les difficultés calculatoires de travailler sur un extrait de ce corpus rassemblant 9 500 paragraphes choisis aléatoirement. La méthode de classification CHAVL, présentée à la section 2.3.3, est utilisée pour réaliser l'analyse des données.

Afin d'effectuer la classification, 9 500 colonnes correspondant à l'extrait aléatoire de paragraphes mentionné et 165 lignes correspondant aux noms apparaissant plus de 8 fois sur cette sélection sont extraites de la matrice $\mathcal{M}$ définie à la section 2.3.1.3, qui rassemble les informations de répartition sur l'ensemble du corpus. Ces données constituent pour CHAVL un « tableau de contingence » : l'objet élémentaire considéré étant l'occurrence de lemme de mot (sous une forme fléchie quelconque), chaque case m_{ij} de la matrice contient le nombre d'objets présentant la double propriété d'être à la fois une forme du lemme i et située dans le paragraphe j. `chavleps` propose une

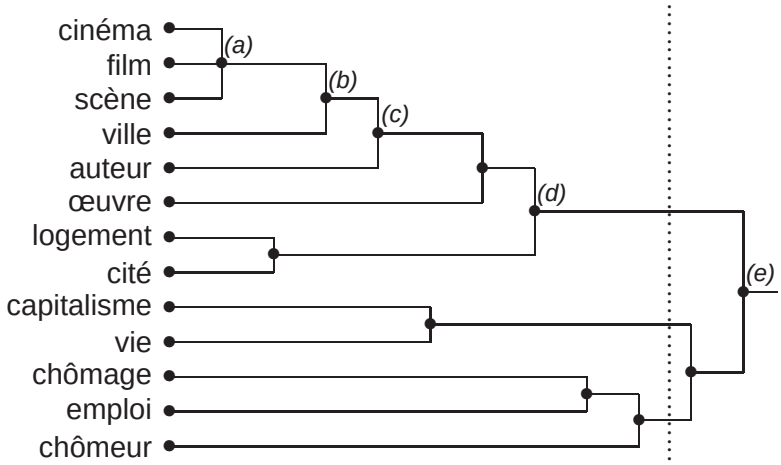


FIG. 2.4 – *Exemple réduit d'arbre de classification des noms produit par* `chavleps`. *La ligne pointillée indique un niveau de lecture « classique » de tels arbres*

mesure de similarité prédéfinie conçue afin de rapprocher les lignes d'un tel tableau de contingence (la formule employée, ainsi que ses justifications statistiques, sont présentées dans [LLP94]). C'est cette mesure qui est utilisée par les auteurs de [PS00] afin de rapprocher les mots selon la similarité de leur répartition sur les paragraphes du corpus.

Le résultat de cette classification est un arbre du type de celui présenté à la figure 2.4, quoique d'une cohérence plus incertaine. Afin d'en extraire des classes thématiques, les auteurs font usage de l'indication faite par CHAVL de « niveaux significatifs » dans l'arbre (voir section 2.3.3), et retiennent toutes les classes dont la formation a lieu à de tels niveaux significatifs (*i.e.* les classes dont le nœud correspondant dans l'arbre apparaît sur l'un de ces niveaux). Ils obtiennent par cette sélection 45 classes, qui sont soumises à une expertise humaine. À l'issue de ce filtrage manuel, 21 classes sont retenues et considérées comme révélatrices d'un thème précis. Ces résultats sont suffisamment positifs pour que l'on puisse considérer le principe fondamental comme valide, mais une intervention humaine reste nécessaire avant d'obtenir un ensemble de classes de mots-clés « propres ». Afin d'améliorer ces résultats et de les rendre compatibles avec un mode de fonctionnement totalement automatique, nous partons donc d'une base similaire concernant la partie « analyse de données » (mesure de similarité entre mots fondée sur la répartition des noms parmi les paragraphes et mode de classification hiérarchique) et travaillons à la définition d'une méthode plus raffinée de lecture de l'arbre de classification obtenu — nous désignons par ce terme de lecture l'opération consistant à analyser l'arbre afin d'en extraire les classes les plus pertinentes pour la tâche à accomplir.

Cette évolution algorithmique nous permet soit d'augmenter « raisonnablement » le nombre de mots considérés par rapport à [PS00] (passant de 165 à

quelque 400) en obtenant un résultat de meilleure qualité, soit de faire croître ce nombre plus « radicalement » et de rester alors à une précision de résultat similaire à ceux présentés ci-dessus. Profitant du surcroît de robustesse apporté au système dans son ensemble par les traitements subséquents, qui permettent de filtrer assez efficacement le « bruit » sur les résultats, nous faisons le second choix, et traitons cette fois les 1 000 noms et adjectifs[8] les plus fréquents d'un extrait aléatoire de 10 000 paragraphes (nous écartons de la sélection de mots étudiée les mots « neutres » aisément repérables tels les noms de mois ou unités de mesure). Nous présentons néanmoins à la section 2.5.3 les résultats obtenus pour une sélection de 400 noms, afin de fournir un élément de comparaison avec [PS00] illustrant l'intérêt des traitements présentés ici.

La méthode d'extraction automatique de classes de mots-clés de l'arbre de classification produit par CHAVL que nous nous proposons de développer doit surmonter deux difficultés principales. D'une part, même dans le cas d'une classification de relativement bonne qualité, comme celle de la figure 2.4, les classes susceptibles de retenir notre intérêt se forment à des hauteurs très variables dans l'arbre de classification ; il n'est donc pas pertinent de faire usage du mode traditionnel de lecture « par niveau » de l'arbre — l'expérience décrite dans [PS00] montre en outre que les « niveaux significatifs » proposés par la méthode CHAVL ne consituent pas un indice suffisamment fiable pour guider totalement cette lecture. D'autre part, les choix de fusions effectués par CHAVL, motivés statistiquement, ne correspondent pas toujours à ce que l'on espérerait d'un point de vue sémantique, à cause notamment des effets de polysémie forte (ou homonymie). Nous désirons donc mettre au point un mode de lecture qui permette non seulement d'extraire automatiquement les classes les plus pertinentes à différents niveaux de l'arbre de classification, mais également d'opérer, si nécessaire, des réorganisations mineures de cet arbre pour éviter l'insertion de quelques « intrus » dans des classes thématiquement homogènes, voire de réinsérer de tels éléments dans une autre classe où leur présence serait plus judicieuse.

La solution proposée pour résoudre ce problème consiste à comparer la classification opérée sur les mots en fonction de leurs paragraphes d'apparition avec sa « réciproque », classification des paragraphes en fonction de la ressemblance des ensembles de mots qu'ils contiennent. Nous présentons ci-dessous les fondements de cette approche, ainsi que sa mise en œuvre.

[8] Nous avons pu constater au cours de nos expériences que les adjectifs constituaient souvent des indicateurs thématiques aussi fiables que les noms. Prendre en compte noms et adjectifs pour notre étude nous permet d'augmenter le nombre de mots considérés à 1 000 sans faire baisser de manière trop importante leur nombre d'occurrences moyen, ce qui est nécessaire à la représentativité des mesures statistiques effectuées.

2.5.2. Exploiter l'arbre de classification des mots grâce à une classification des paragraphes

Nous procédons dans ce qui suit à un regroupement des paragraphes du corpus d'étude extrait en fonction de leur similarité lexicale, c'est-à-dire de la ressemblance du vocabulaire qu'ils développent, en faisant l'hypothèse que cette proximité de vocabulaire implique une proximité thématique — suivant ainsi le principe de « cohésion textuelle » des travaux [Hea94, SSBM96] présentés à la section 2.2. Cette classification est ensuite mise à profit pour développer une méthode de recherche de classes optimales dans l'arbre de classification des mots, en confrontant la qualification thématique opérée par les classes de mots potentielles à la partition des paragraphes en classes thématiques, utilisée comme référent, grâce à une mesure évaluant la corrélation existant entre ces deux réponses apportées au même problème de la reconnaissance de thèmes.

Il est fait mention à de nombreuses reprises dans ce chapitre de classifications effectuées, d'une part, sur des mots, d'autre part, sur des paragraphes. Afin d'éviter aussi bien les ambiguïtés que les périphrases pesantes, nous employons le préfixe « m- » pour distinguer les concepts se référant aux mots (« m-classe » désigne ainsi une classe de mots), le préfixe « p- » étant pour sa part réservé aux paragraphes (« p-classification » : classification effectuée sur les paragraphes).

2.5.2.1. Classification des paragraphes

Cette nouvelle classification nous impose de remettre en question les outils d'analyse statistique employés jusqu'alors. Elle ne peut en effet être effectuée grâce à la même méthode que la m-classification des noms et adjectifs : étant donné le faible remplissage du tableau de contingence — la matrice $\mathcal{M}$ — utilisé (au moins 99 % de ses cases ont une valeur nulle), 1 000 mots sont évidemment insuffisants pour caractériser efficacement 10 000 paragraphes. Il est donc nécessaire d'exploiter une partie aussi importante que possible de $\mathcal{M}$, et de retenir également pour la comparaison les mots peu fréquents. Cela a naturellement pour effet de rendre le tableau de contingence employé pour la p-classification encore plus « creux », à tel point que les mesures statistiques sophistiquées employées par CHAVL pour calculer les similarités entre objets atteignent leurs limites. Nous définissons donc pour rapprocher les paragraphes déployant un vocabulaire similaire un indice de proximité lexicale moins « sensible » d'un point de vue calculatoire, permettant ainsi l'emploi d'une plus grande quantité de données. Cet indice est inspiré de la mesure de cohésion lexicale évoquée en section 2.2, par exemple dans les travaux de M. Hearst ou G. Salton *et al.*, qui consiste à déterminer le nombre de mots partagés par deux paragraphes pour évaluer leur proximité thématique. Nous affinons toutefois ce principe en donnant un poids supplémentaire au partage de

mots rares, dont l'apparition vraisemblable dans peu de thèmes distincts fait de leur coprésence dans deux paragraphes un indicateur fort pour la classification thématique. L'importance de chaque mot est donc inversement proportionnelle à son nombre d'occurrences dans le sous-corpus aléatoire de 10 000 paragraphes sur lequel nous travaillons, et la mesure est normalisée en fonction de la taille des paragraphes comparés. Par conséquent, la similarité entre deux paragraphes A et B est définie par :

$$s\,(A,B) = \frac{1}{\min(n_A, n_B)} \sum_i \frac{\min(a_i, b_i)}{n_i} \tag{2.1}$$

où $A = (a_i)$ et $B = (b_i)$ sont les vecteurs rassemblant le nombre d'occurrences de chaque mot considéré pour ce calcul dans chacun des paragraphes, n_i est le nombre total d'occurrences du mot i dans les 10 000 paragraphes du corpus extrait, et n_A et n_B les nombres de mots des deux paragraphes.

De par sa simplicité, cette mesure est à même de prendre en compte tous les mots du sous-corpus aléatoire sans contrainte de fréquence, et nous permet d'effectuer le calcul pour les noms et adjectifs apparaissant au moins 2 fois[9] dans l'extrait étudié, soit 9 000 mots environ. Ce grand volume de données exploitées permet de compenser partiellement la relative naïveté de s. La demi-matrice $10\,000 \times 10\,000$ contenant les indices de similarité entre paires de paragraphes est utilisée par CHAVL pour construire un arbre de p-classification. Celui-ci est assez bien équilibré, ce qui est en général révélateur d'un déroulement « harmonieux » du processus d'agrégations progressives de la CAH, où les objets à classer se répartissent régulièrement dans l'espace de représentation et forment des classes nettement distinctes et bien caractérisées.

Nous ne pouvons, au risque d'entrer dans un cercle vicieux inextricable, aborder la problématique qui consisterait à rechercher une fois de plus un mode de lecture « intelligente » de l'arbre de p-classification. En outre, celui-ci n'est conçu que comme outil pour l'analyse de l'arbre de m-classification, et notre objectif n'est pas de produire des classes rassemblant *tous* les paragraphes abordant un thème donné mais seulement des classes dont on puisse être relativement sûr que tous leurs paragraphes évoquent des thèmes similaires. Nous extrayons donc simplement par coupure à un niveau où la taille moyenne des classes atteint une valeur que nous avons fixée à douze[10], un ensemble d'environ 600 p-classes (les classes présentant au niveau choisi une taille supérieure au double ou inférieure à la moitié de la taille moyenne choisie sont écartées automatiquement).

Confrontés à l'impossibilité pratique de vérifier manuellement la pertinence d'un arbre de classification rassemblant 10 000 éléments, nous ne

[9] Valeur minimale pour que le mot puisse être présent dans deux paragraphes distincts.

[10] Cette valeur a été déterminée empiriquement, et constitue un bon compromis entre généralisation et cohérence thématique des p-classes.

disposons pour juger de la qualité de cette p-partition que d'indicateurs partiels ou indirects : le relativement bon équilibre de l'arbre obtenu est ainsi en règle générale, comme nous l'avons déjà précisé, le signe d'un déroulement satisfaisant du processus de classification, et des validations manuelles effectuées afin de contrôler la cohérence de quelques p-classes choisies au hasard montrent que celles-ci présentent un degré acceptable de cohérence thématique (en moyenne, plus de 70 % des paragraphes d'une p-classe évoquent un même thème). On ne saurait néanmoins considérer le résultat de cette p-classification comme un résultat satisfaisant en soi pour la détection de thèmes, pour plusieurs raisons qui dépassent le simple problème de qualité : ainsi la procédure décrite est-elle, pour des raisons calculatoires, difficilement applicable à des corpus dépassant 10 000 paragraphes ; de plus, la méthode décrite décrite n'a pour objectif, comme nous l'avons mentionné, que la cohérence des classes, pas leur complétude ; enfin et plus fondamentalement, nous n'avons à l'issue de ce traitement aucune indication permettant de caractériser les thèmes justifiant les regroupements en classes effectués.

Malgré toutes ces réserves, la p-classification rudimentaire obtenue constitue un bon « point d'appui » pour diriger l'extraction des m-classes de l'arbre de m-classification, comme nous le présentons maintenant.

2.5.2.2. Définition d'un critère numérique de qualité d'une m-classe par comparaison avec la p-classification

La p-classification est utilisée pour définir une mesure de qualité des classes de mots et nous guider dans le choix des m-classes proposées par l'arbre de m-classification, en apportant à celles-ci de légères modifications si nécessaire. Afin de fonder cette mesure, nous revenons à la « finalité » des m-classes.

Les listes de mots porteuses de thèmes que nous voulons obtenir ont pour objectif d'être utilisées d'une façon que nous avons déjà définie, à savoir : si deux mots d'une m-classe caractérisant un thème sont présents dans un paragraphe du corpus, alors le paragraphe évoque ce thème — on dira par la suite que la m-classe *reconnaît* le paragraphe. Par conséquent, l'ensemble des m-classes que nous voulons extraire de l'arbre de m-classification effectue une classification thématique des paragraphes étudiés, ce qui est aussi le rôle de la p-partition que nous venons de définir. Ces deux classifications devraient donc coïncider autant que possible[11]. Si nous considérons le cas idéal où tous les paragraphes d'une p-classe abordent un même thème et où chaque m-classe reconnaît l'ensemble des paragraphes évoquant le thème qu'elle caractérise, et seulement eux, alors une m-classe reconnaît soit tous les paragraphes d'une p-classe, soit aucun. La mesure de qualité définie donne la préférence aux m-classes les plus proches de cette configuration idéale.

[11]La correspondance est essentiellement limitée par le fait qu'un paragraphe peut être reconnu par aucune, une ou plusieurs m-classes, mais appartient à une et une seule p-classe.

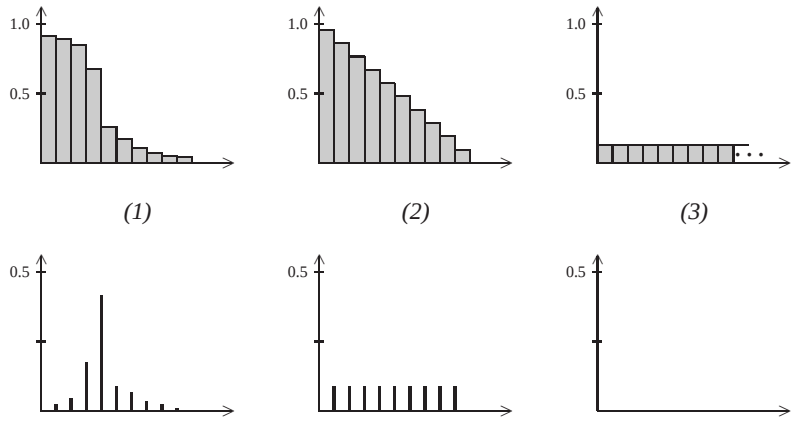


FIG. 2.5 – *Configurations typiques : en haut, trois graphiques montrant, pour une collection de p-classes (axe des abscisses) quelle proportion de leurs paragraphes est reconnue par une m-classe donnée (axe des ordonnées) ; en bas, les « dérivées » de ces diagrammes*

Définition mathématique d'un critère numérique de qualité *q* Soit $\mathcal{C}$ une m-classe, et $\mathcal{P}_1, \ldots, \mathcal{P}_p$ l'ensemble des p-classes définies par la p-partition obtenue en 2.5.2.1 (où $p = 600$). Nous écrirons dans la suite $\mathrm{rec}\,(\mathcal{C}, P)$ pour exprimer le fait que $\mathcal{C}$ reconnaît un paragraphe P donné. À $\mathcal{C}$, nous associons le vecteur $\overrightarrow{C} \in \mathbb{R}^p$, défini par :

$$\overrightarrow{C} = (c_1, \ldots, c_p)\,,\ \text{où}\ \forall i \in [1,p]\,, c_i = \frac{\mathrm{Card}\,\{P \in \mathcal{P}_i \,|\, \mathrm{rec}\,(\mathcal{C}, P)\}}{\mathrm{Card}\,(\mathcal{P}_i)} \qquad (2.2)$$

Chaque élément c_i de $\overrightarrow{C}$ est la proportion de paragraphes de $\mathcal{P}_i$ reconnus par $\mathcal{C}$. Nous définissons $\overrightarrow{C}' \in \mathbb{R}^p$, $\overrightarrow{C}' = \left(c'_1, \ldots, c'_p\right)$, vecteur rassemblant les mêmes valeurs que $\overrightarrow{C}$ mais classées par ordre décroissant — la numérotation des $\mathcal{P}_i$ étant totalement arbitraire, l'ordre dans lequel sont présentés les c_i n'a aucune importance intrinsèque. C'est de l'étude du profil général de ce vecteur que nous dérivons notre mesure de qualité.

La figure 2.5 présente de manière schématique plusieurs configurations typiques de ce profil. La première s'approche clairement du cas idéal recherché, où toutes les proportions valent 1 ou 0. Dans la seconde, des différences entre les taux de reconnaissance des diverses p-classes sont toujours notables, mais il n'existe pas de ligne de démarcation claire entre deux catégories. Le dernier cas, enfin, est le moins favorable, et correspond à une situation où la m-classe n'exprime aucune préférence pour l'une ou l'autre des p-classes.

Afin de détecter les vecteurs $\overrightarrow{C}'$ présentant, comme le premier exemple, une dichotomie claire entre deux ensembles de p-classes, nous pourrions faire usage d'une simple mesure de l'écart type de $\overrightarrow{C}'$. Malheureusement, l'écart type atteint son maximum lorsque la moitié exactement des valeurs consi-

dérées vaut 1, et l'autre 0, ce qui aurait pour effet de favoriser des classes de mots reconnaissant la moitié des paragraphes du corpus — nous ne pouvons pas avancer *a priori* de valeur pour le nombre de paragraphes évoquant un thème donné, mais 50 % du corpus semble sans conteste excessif. Afin de mettre au point une mesure ne reflétant que la « brutalité » du passage entre valeurs fortes et faibles, indépendamment des tailles de ces deux ensembles, nous définissons un nouveau vecteur $\overrightarrow{C}'' \in \mathbb{R}^{p-1}$ par :

$$\overrightarrow{C}'' = \left(c''_1, \ldots, c''_{p-1}\right), \text{ où } \forall i \in [1, p-1], c''_i = c'_i - c'_{i+1} \tag{2.3}$$

$\overrightarrow{C}''$ rassemble la séquence de toutes les différences entre valeurs consécutives de $\overrightarrow{C}'$, représentant pour ainsi dire sa « dérivée » (pour chaque exemple de $\overrightarrow{C}'$ représenté sur la figure 2.5, $\overrightarrow{C}''$ est représenté sur la ligne inférieure).

Une valeur importante de l'écart type de $\overrightarrow{C}''$ (noté $\sigma_{\overrightarrow{C}''}$) reflète l'existence d'un pic dans cette « dérivée », et donc d'un point de rupture important entre valeurs faibles et fortes de $\overrightarrow{C}'$. $\sigma_{\overrightarrow{C}''}$ nous permet ainsi d'établir une distinction numérique entre les profils (1) et (2) de la figure 2.5, mais est nul pour (2) aussi bien que (3). Afin d'étendre les capacités de discrimination de notre mesure à ce type de cas, on combine $\sigma_{\overrightarrow{C}''}$ avec l'extension de $\overrightarrow{C}'$ (*i.e.* la valeur $\left(c'_1 - c'_p\right)$), obtenant ainsi l'expression suivante pour la « fonction de qualité » q :

$$q\left(\mathcal{C}\right) = \left(1 + \sigma_{\overrightarrow{C}''}\right)\left(1 + \left(c'_1 - c'_p\right)\right) - 1 \tag{2.4}$$

Le choix d'une formule du type $(1+a)(1+b)-1$ nous permet d'équilibrer assez simplement les contributions respectives des deux critères considérés (écart type de $\overrightarrow{C}''$ et extension de $\overrightarrow{C}'$). Ainsi définie, q reflète assez fidèlement la correspondance existant entre la classification des paragraphes résultant de l'usage de $\mathcal{C}$ et la p-partition obtenue précédemment. Pour les trois exemples de profils donnés à la figure 2.5, les valeurs obtenues pour q sont environ $1,25$, $0,95$ et 0, respectivement. Il est important de noter que q n'est pas une mesure de similarité, mais une évaluation *a posteriori* de la qualité d'une classe, ce qui nous empêche de l'employer directement pour rassembler des mots-clés (la seule méthode envisageable pour cela consisterait à réaliser une recherche exhaustive dans l'espace de toutes les partitions en classes envisageables pour l'ensemble des mots-clés potentiels, ce qui est bien entendu d'une complexité insurmontable) ; elle permet en revanche de porter un « jugement » sur une classification suggérée, capacité que nous mettons maintenant à profit dans le cadre d'un algorithme de lecture de l'arbre de m-classification.

2.5.2.3. Algorithme de lecture de l'arbre de m-classification exploitant la mesure de qualité q

L'intérêt de la mesure q pour la lecture de l'arbre de m-classification est double : d'une part, elle sert à pointer les classes pertinentes dans l'arbre quel que soit leur niveau ; d'autre part, elle permet d'ignorer certains regroupements de mots effectués par CHAVL, voire de les modifier. L'algorithme développé à cette fin est présenté formellement en Algorithme 1, page 67. De manière plus descriptive son fonctionnement est le suivant : il part des feuilles de l'arbre de m-classification et remonte vers la racine en vérifiant à chaque nœud rencontré si la fusion proposée accroît la valeur de q. Si c'est le cas, l'algorithme accepte la fusion et continue l'exploration ascendante de l'arbre avec cette nouvelle classe. Sinon, l'algorithme continue à remonter vers la racine mais sans prendre en compte la fusion. Dans ce cas, la remontée de l'arbre se poursuit avec un ensemble de classes au lieu d'une simple m-classe. La figure 2.6 présente étape par étape l'évolution du processus de lecture de l'arbre présenté à la figure 2.4, page 58, en partant des classes « élémentaires » de premier niveau, dont l'algorithme réalise dans tous les cas la fusion (sauf cas réellement pathologique, la qualité q d'une classe de mots est toujours supérieure à celle d'un mot seul, et les fusions « de première génération » sont donc toujours considérées comme valides par l'algorithme de lecture). Au nœud (b) de la figure 2.4 (étape 1), par exemple, le programme décide de ne pas effectuer la fusion et poursuit l'exploration ascendante de l'arbre avec un ensemble de deux classes : {{*cinéma, film, scène*}, {*ville*}}.

Aux nœuds supérieurs, toutes les possibilités de fusion entre membres des ensembles de classes sont testées pour détecter les plus intéressantes en termes d'évolution de q. On aura par exemple au nœud (c) (étape 2) : {{*cinéma, film, scène*}, {*ville*}} « + » {*auteur*} → {{*cinéma, film, scène, auteur*}, {*ville*}}

Cette méthode peut permettre de réaffecter un mot « laissé pour compte » à une autre classe, comme c'est le cas à l'étape 4 : *ville* est agrégé à la classe {*logement, cité*}, regroupement « manqué » par CHAVL.

Finalement, nous obtenons en (e), racine de l'arbre, la partition : {{*cinéma, film, scène, auteur, œuvre*}, {*ville, logement, cité*}, {*capitalisme,vie*}, {*chômeur, emploi, chômage*}}.

Les classes extraites sont donc différentes de celles produites par CHAVL, et sont obtenues à l'issue du parcours sans qu'il soit nécessaire de les chercher à divers niveaux de l'arbre. Une heuristique consistant à détecter les classes « stabilisées » (c'est-à-dire peu susceptibles d'améliorer leur qualité par fusion avec une autre classe) et à les écarter du processus de recherche des meilleures fusions permet de réduire le temps de calcul[12].

[12] L'ensemble du système de construction / lecture de l'arbre de m-classification s'exécute en 5 à 10 minutes sur un ordinateur personnel « moderne », à base par exemple de Pentium IV.

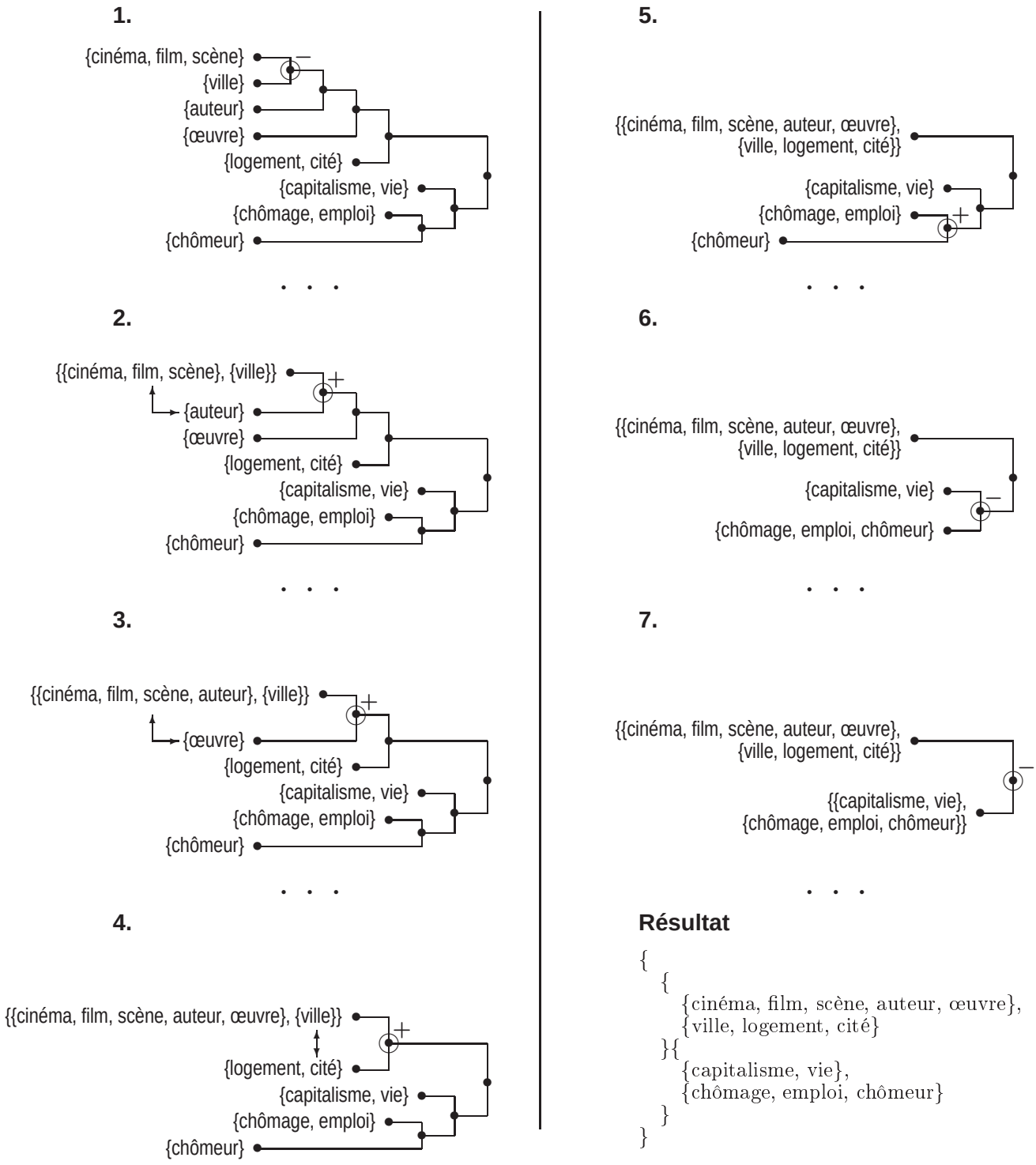


FIG. 2.6 – *Déroulement de l'algorithme d'extraction de m-classes. Le nœud de l'arbre traité à une étape donnée est marqué par un cercle ; « − » indique que la fusion proposée n'est pas pertinente au regard de l'évolution de la qualité q, « + » qu'une fusion est possible. Dans ce cas, si plusieurs agrégations sont envisageables, une double flèche indique la meilleure*

ALGORITHME 1 *Expression récursive formelle de l'algorithme de lecture d'un arbre de classification des mots exploitant la mesure de qualité q*

fonction extractionPartition - **paramètre** : N, nœud de l'arbre
  Si N est une feuille, **résultat** : $\{\{valeur(N)\}\}$
  classes = $\{\bigcup_{i=1..nbFils} \{C_i \mid C_i \in (\text{extractionPartition}\,(fils_i) \cup \{\varnothing\})\}\}$
  partition = $\varnothing$
  répéter
    C = **choisir** $c \in$ classes **telle que** $q(c) = \max_{d \in \text{classes}}(q(d))$
    partition = partition $\oplus$ C
    classes = classes $\setminus \{c \in$ classes $\mid c \cap C \neq \varnothing\}$
  jusqu'à classes = $\{\varnothing\}$
résultat : partition

2.5.3. Principe de l'étape suivante

Cette première étape de construction de classes sémantiques est présentée dans [RS02]. Nous y montrons qu'employée isolément sur des données du même ordre de grandeur que celles de [PS00] (quelque 400 noms et un extrait de 8 000 paragraphes), elle permet une amélioration significative par rapport aux résultats de ce travail : elle propose en moyenne (les résultats étant amenés à varier suivant l'extrait aléatoire de paragraphes choisi) un total de 35 classes parmi lesquelles 25 peuvent être considérées comme valides, c'est-à-dire révélatrices sans ambiguïté d'un thème clairement identifiable, tous leurs mots étant liés à ce thème. Nous passons donc d'une précision de 47 % chez [PS00] (21 classes sur 45) à environ 71 %. Parmi les « bonnes » classes obtenues, on peut par exemple noter celle-ci, représentative du thème « enseignement / recherche » : {*bureau*, *centre*, *enseignement*, *institution*, *recherche*, *université*, *école*}.

Nous ne nous étendrons pas sur l'analyse de ces résultats, toujours intermédiaires, nous bornant à rappeler deux sources d'insuffisances intrinsèques limitant nécessairement la précision de la méthode employée jusqu'ici, et déjà évoquées en introduction :

- le mode de sélection des mots étudiés, guidé par des exigences statistiques de nombre d'occurrences, ne nous offre aucune garantie que ces mots aient réellement vocation à jouer le rôle de mots-clés ;
- le fait de ne travailler que sur un extrait du corpus rend les résultats tributaires de la représentativité de l'extrait choisi.

Le second de ces problèmes est à l'origine de la solution implantée dans FAESTOS permettant de les surmonter : il implique en effet que les résultats obtenus sont amenés à varier en fonction de l'extrait choisi. Un intérêt de cette variabilité est que les rassemblements de mots que l'on peut considérer comme

du « bruit » — en particulier, les regroupements mettant en jeu des mots sans coloration thématique notable — se reproduisent rarement à l'identique (par définition même du bruit), alors que les regroupements « légitimes » apparaissent assez fréquemment. Pour parler en termes probabilistes : plus deux mots sont effectivement fortement liés et révélateurs d'un thème, plus ils seront regroupés souvent si l'on considère n exécutions successives du système décrit, réalisées sur des extraits différents du corpus. Pour n suffisamment grand, on peut en outre considérer qu'une très grande majorité des données textuelles disponibles a été prise en compte au moins une fois au cours des classifications, et qu'on a ainsi « simulé », à la façon d'un échantillonnage statistique de type « Monte Carlo » [ET91], une étude de l'ensemble du corpus. Ce principe constitue le fondement de la seconde étape de traitement, présentée en section suivante, qui permet d'obtenir un premier ensemble de m-classes restreintes mais exemptes d'« intrus », classes que nous nommons *noyaux*.

2.6. Obtention de noyaux thématiques

Conformément aux considérations exposées précédemment, la problématique à résoudre pour cette deuxième étape du système FAESTOS consiste à construire de manière automatique, à partir d'un ensemble de m-classifications obtenues au cours de n exécutions du traitement présenté à la section précédente[13], une classification telle que les regroupements de mots qu'elle réalise correspondent aux associations les plus couramment observées dans les n m-classifications. L'algorithme que nous avons mis au point à cette fin s'appuie sur le formalisme des graphes.

Les n partitions obtenues sont synthétisées sous la forme d'un graphe valué $G = (X, \Gamma, v)$, où X est l'ensemble de tous les mots classés au moins une fois, Γ l'ensemble des paires de mots réunies dans une même classe au moins une fois dans l'une des n partitions, et la fonction de poids $v : \Gamma \rightarrow \mathbb{N}$ associe à chacune de ces paires le nombre de fois où elles se sont trouvées ainsi rassemblées. Une partie du graphe défini de cette manière est représenté sur la figure 2.7, où trois groupes « denses » de mots peuvent être observés autour des axes forts « *université* / *étudiant* », « *santé* / *soin* » et « *recherche* / *scientifique* ». Chacune des classifications prises en compte pour la construction de ce graphe ayant réalisé une sélection des 1 000 noms et adjectifs les plus fréquents sur des sous-corpus aléatoires distincts, des variations sont observables dans ces sélections ; X rassemble donc sensiblement plus de 1 000 sommets — environ 1 100 en moyenne si $n = 45$, ce qui montre que malgré ces variations un important « tronc commun » de mots a été considéré dans une grande majorité des m-classifications considérées.

[13] Expérimentalement, nous avons constaté que la valeur idéale pour n s'établit autour de 45. Nous n'avons pu définir s'il existait un lien formalisable entre, par exemple, la taille du corpus ou des sous-corpus aléatoires extraits et cette valeur.

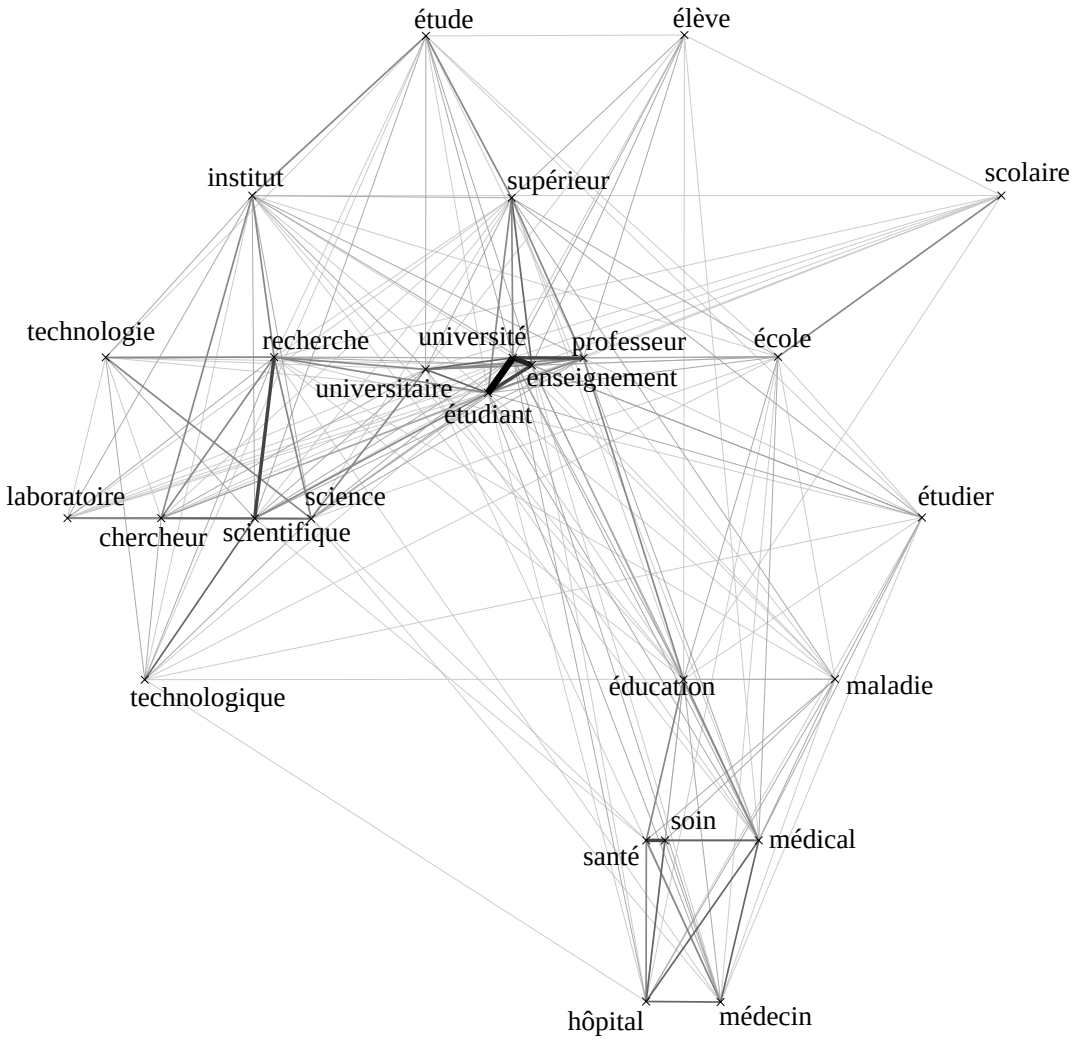


FIG. 2.7 – *Extrait du graphe de regroupement des mots-clés potentiels construit à partir de* 45 *classifications réalisées sur des sous-corpus aléatoires distincts de 10 000 paragraphes, chacune prenant en compte les 1 000 noms et adjectifs les plus fréquents sur cet extrait. Les mots représentés sont tous ceux connectés à « université » dans le graphe. Leurs positions ont été calculées par un algorithme tentant de rapprocher d'autant plus deux sommets que l'arc les reliant a un poids important (arc épais et foncé sur la figure). Par souci de lisibilité, les arcs de poids 1 n'ont pas été représentés. Trois groupes se détachent autour des axes « université / étudiant », « santé / soin » et « recherche / scientifique »*

ALGORITHME 2 *Extraction de classes de sommets fortement connexes à partir du graphe synthétisant les résultats de m-classifications multiples*

fonction classesDeGraphe - **paramètre** : un graphe valué $G = (X, \Gamma, v)$
classes = $\varnothing$
répéter
choisir $(a, b) \in \Gamma$ t.q. $V = v(a, b) = \max \{v(\alpha, \beta) \mid (\alpha, \beta) \in \Gamma\}$
à chaque $V' \leq V$, associer $\mathcal{G}_{V'}$ composé des sommets n t.q. :
(1) $\forall\, n' \in X$, si $(n, n') \in \Gamma$ et $v(n, n') \geq V'$, alors $n' \in \mathcal{G}_{V'}$
(2) $\exists\, m_0, ..., m_i \in X$ tels que
$m_0 \in \{a, b\}$ et $m_i = n$
$\forall\, j < i,\ v(m_j, m_{j+1}) \leq v(m_{j-1}, m_j)$ et $v(m_j, m_{j+1}) \geq V'$
choisir $W = \min \{V' \leq V \mid \mathcal{G}_{V'} \neq \varnothing \text{ and Card}(\mathcal{G}_{V'}) \leq 11\}$
si W non défini **alors interrompre itération**
classes = classes $\oplus\, \mathcal{G}_W$
retirer de G tous les sommets de $\mathcal{G}_W$
jusqu'à $\forall\, (a, b) \in \Gamma,\ v(a, b) = 1$
résultat classes

L'extraction des « noyaux » thématiques consiste à rechercher des parties fortement connexes du graphe complet, et est réalisée par l'Algorithme 2. Le principe de cet algorithme consiste à sélectionner l'arc de poids le plus fort[14] dans G (que nous désignons désormais comme l'arc (a, b), de poids V), et de définir de manière automatique un seuil au-dessous duquel tous les arcs sont ignorés. Une fois ce seuil défini, les sommets appartenant à la composante connexe du graphe contenant a et b sont extraits pour former un noyau thématique ; ils sont ôtés du graphe et la procédure est répétée jusqu'à ce qu'aucun nouveau noyau ne puisse être extrait (condition précisée ci-dessous).

La principale difficulté réside dans la définition d'un seuil d'activation des arcs permettant la génération de classes consistantes de tailles convenables. Pour cela, nous définissons la famille des sous-graphes $\mathcal{G}_{V',\, 1 \leq V' \leq V}$, qui pour chaque valeur de V' rassemblent les sommets qui peuvent être atteints à partir de a ou b par un chemin composé exclusivement d'arcs dont le poids est au moins V' — il s'agit en fait du sous-graphe connexe subsistant autour de a et b si l'on supprime tous les arcs de poids inférieur à V'. Si tout sommet de ce sous-graphe $\mathcal{G}_{V'}$ peut être atteint à partir de a ou b par une succession d'arcs de poids monotonement décroissants, alors V' est un seuil potentiel valide, sinon on considère qu'il rassemble autour de l'axe (a, b) des sommets appartenant à d'autres sous-composantes logiques de G. La figure 2.8 montre l'exemple d'un graphe « jouet » défini afin d'illustrer cette notion. Le seuil de coupe choisi

[14] Au cas où plusieurs arcs présentent un poids maximal égal, la sélection entre eux est aléatoire ; les résultats obtenus sont indépendants de l'ordre dans lequel ils sont considérés.

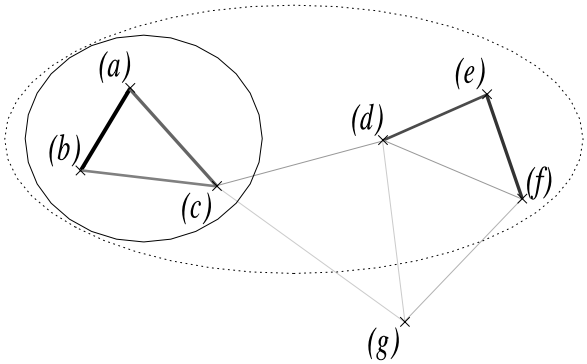


FIG. 2.8 – *Exemple de graphe simple illustrant le processus de sélection de seuil de coupe pour la construction de noyaux thématiques :* (a, b) *étant l'arc de poids le plus fort du graphe, la ligne continue délimite l'ensemble des sommets sélectionnés en employant un seuil de coupe relativement élevé, tandis que la ligne pointillée délimite ceux sélectionnés par un seuil plus bas (comme sur la figure 2.7, les arcs ont un poids d'autant plus important qu'ils sont foncés et épais). Le second seuil ne constitue pas selon notre critère un choix valide, car aucun chemin d'arcs de poids décroissants ne mène de* (a) *ou* (b) *à* (e)

par l'algorithme de recherche de composantes fortement connexes est en général le plus petit seuil valide, à moins que celui-ci ne génère un sous-graphe d'une taille trop importante (ce qui peut se produire vers la fin de l'exécution de l'algorithme, lorsque restent dans G de nombreux sommets faiblement connectés) ; dans ce cas, le plus petit seuil V' générant un sous-graphe dont le nombre de sommets est inférieur à 1 % du nombre initial de sommets de G est retenu (dans le cas de nos expériences, X rassemblant environ 1 100 mots, cela correspond à une taille limite de 11 sommets). S'il n'est pas possible de trouver un tel seuil (autrement dit, si V lui-même génère un sous-graphe connexe de taille trop importante), l'algorithme est interrompu et les sommets restants sont écartés de la construction des noyaux thématiques. Au cours de toutes les applications de l'algorithme que nous avons pu réaliser, c'est cette condition qui entraîne l'arrêt de l'itération. Une condition « garde-fou » est néanmoins spécifiée afin de garantir la terminaison du programme : celui-ci interrompt sa recherche lorsque tous les arcs restant dans G sont de poids 1, autrement dit lorsque tous les liens restant entre mots peuvent correspondre à du « bruit ».

Nous obtenons grâce à cet algorithme de 40 à 45[15] noyaux thématiques, ou « prototypes de classes », rassemblant chacun de 5 à 10 mots. Il ressort de ces chiffres que plus de la moitié des 1 100 noms et adjectifs présents dans les classifications obtenues par la première étape ne sont présents dans aucun des noyaux construits ; c'est au prix d'une telle « sévérité » de filtrage que l'algo-

[15]Bien que la mise en commun de multiples exécutions du premier système de construction de m-classes permettent une importante stabilisation des résultats, de légères variations sont encore observables. C'est pourquoi nous ne donnons toujours que des valeurs moyennes.

rithme s'avère suffisamment robuste pour supporter une quantité de bruit assez importante sur les classifications prises en compte, ce qui rend possible l'extension de l'étude à 1000 noms et adjectifs (plutôt que 400 noms). Les noyaux obtenus comprennent ainsi des classes telles que :

- {*école*, *élève*, *enseignement*, *étudiant*, *professeur*, *universitaire*} (thème « enseignement »),
- {*chercheur*, *laboratoire*, *recherche*, *science*, *technique*, *technologie*} (thème « recherche »).

Ces résultats montrent bien l'intérêt du traitement effectué : d'une part, en rendant possible l'extension de la liste des mots-clés potentiels et en sophistiquant leur regroupement, il permet de raffiner la granularité de l'analyse pour les domaines fréquemment abordés (la classe présentée à la section précédente regroupait enseignement et recherche en un seul thème) ; d'autre part, il réalise bien comme souhaité un filtrage des mots classés et ne conserve que ceux qui sont fortement associés à un thème (par rapport à la m-classe obtenue précédemment, les mots *bureau*, *centre* et *institution* ont disparu).

Toutefois, les mots composant nos noyaux représentent toujours un sous-ensemble de ceux sélectionnés sur un critère de fréquence lors de la première étape de construction de classes. Nous ne disposons en conséquence à l'issue de cette étape de calcul que de mots à la fois fortement « colorés » thématiquement et apparaissant à de nombreuses reprises dans le corpus. Un dernier traitement nous permet maintenant d'étendre les noyaux obtenus afin d'y adjoindre les mots-clés de fréquence moindre.

2.7. Extension des noyaux de classes de mots-clés

Il est possible à ce stade du calcul de faire appel à un algorithme d'apprentissage de type supervisé : en effet, la collection de noyaux thématiques extraits constitue un premier ensemble de classes rudimentaires, et donc une première réponse au problème posé de caractérisation et détection de thèmes. Nous disposons ainsi d'un recensement des thèmes qu'il nous sera possible d'identifier et, pour chacun d'eux, d'un critère de détection de leur présence, par cooccurrence de mots-clés (étant donnée la taille limitée des noyaux, ce critère ne permet de reconnaître qu'un nombre de paragraphes assez faible, mais néanmoins suffisant en pratique pour l'apprentissage). Cette troisième et dernière partie du système FAESTOS, dont l'objectif est d'accroître le nombre de termes rassemblés dans chaque classe en tirant profit des mots de plus faible nombre d'occurrences, fonctionne donc selon un principe simple : ajouter à chacun des noyaux les mots présentant sur les paragraphes qu'il reconnaît une fréquence anormalement élevée par rapport à leur fréquence moyenne sur l'ensemble du corpus — signe que ces mots possèdent un lien particulier avec le thème caractérisé par le noyau. Étant donnée la légèreté algorithmique de la procédure mise en œuvre, nous pouvons maintenant travailler sur l'intégralité des

ALGORITHME 3 *Expansion des prototypes de classes de mots-clés*

fonction expClasses - **paramètre** : $\{\mathcal{C}_{1...n}\}$ un ensemble de noyaux
soit LP_i la liste des paragraphes reconnus par $\mathcal{C}_i$
répéter
pour chaque $\mathcal{C}_i$, pour chaque mot-clé potentiel M_j, calculer

$$r_{ij} = \frac{\text{fréquence de } M_j \text{ dans } LP_i}{\text{fréquence de } M_j \text{ dans le corpus}} \tag{2.5}$$

trouver I_{max} et J_{max} t.q. $r_{I_{max}J_{max}} = \max_{i,j}(r_{ij})$
$\mathcal{C}_{I_{max}} = \mathcal{C}_{I_{max}} \oplus M_{J_{max}}$
mettre à jour $LP_{I_{max}}$
tant que l'ajout de $M_{J_{max}}$ à $\mathcal{C}_{I_{max}}$ augmente le nombre moyen de mots de $\mathcal{C}_{I_{max}}$ par paragraphe de $LP_{I_{max}}$
résultat $\{\mathcal{C}_{1...n}\}$

quelque 98 000 paragraphes du corpus, en considérant comme mots-clés potentiels les 3 600 noms et adjectifs y apparaissant le plus fréquemment, c'est-à-dire plus de 100 fois[16].

À chacun des prototypes de classes $\mathcal{C}_i$ obtenus à l'étape précédente, on associe la liste LP_i des paragraphes qu'il reconnaît, puis on répète le traitement présenté à l'Algorithme 3 : la mesure employée pour mesurer le lien existant entre la classe de mots-clés en construction $\mathcal{C}_i$ et le mot M_j, notée r_{ij}, est le rapport entre la fréquence d'apparition de M_j dans les paragraphes reconnus par $\mathcal{C}_i$ comme abordant « son » thème et la fréquence moyenne de M_j dans la totalité du corpus. C'est une mesure volontairement simple, assez proche du très classique $tf \times idf$ employé en indexation de documents.

Au fur et à mesure de la croissance des classes par ajout des mots présentant avec elles les plus fortes valeurs de r_{ij}, les listes de paragraphes LP_i (paragraphes détectés par les $\mathcal{C}_i$ comme abordant leur thème) sont recalculées afin de refléter l'extension des capacités de détection des $\mathcal{C}_i$. L'intérêt de cette mise à jour se comprend aisément en termes statistiques : nous souhaitons calculer un facteur de fréquence d'apparition relative pour un mot M_j entre les textes évoquant un thème donné T et l'ensemble du corpus. Si $\mathcal{C}_i$ est la classe associée au thème T, l'indice r_{ij} est une valeur approchée de ce facteur, calculée en approximant l'ensemble des paragraphes abordant T par l'ensemble des paragraphes reconnus par $\mathcal{C}_i$ à une étape donnée de l'algorithme. À mesure que l'algorithme progresse et que la taille des $\mathcal{C}_i$ et LP_i augmente, cette approximation est de plus en plus précise, et donc la valeur de r_{ij} reflète de

[16] Prendre en compte des mots moins fréquents permet d'augmenter le nombre de mots finalement classés, mais ne présente qu'un intérêt très marginal en termes de qualité de la détection de thèmes.

plus en plus finement l'association de M_j avec les textes abordant T. Ainsi, lors des premières itérations de l'algorithme, les termes sélectionnés présentent des valeurs de r_{ij} très fortes (jusqu'à 200), ce qui permet d'ignorer la grande amplitude de l'intervalle de confiance (ou moins formellement, de l'incertitude) sur cette valeur induit par la petite taille des LP_i (quelques dizaines de paragraphes). À mesure que la construction des classes de mots-clés progresse, les valeurs de r_{ij} considérées sont de plus en plus faibles, mais cela est compensé par la réduction de l'amplitude de leur intervalle de confiance (due à la croissance des LP_i).

Nous interrompons la croissance des classes dès qu'un ajout aurait pour effet de faire diminuer le nombre moyen de mots-clés par paragraphe reconnu. Ceci se justifie par le fait qu'une baisse de cette densité moyenne correspond le plus souvent à l'ajout d'un mot tendant à étendre le thème original vers une notion annexe : son association avec les mots déjà présents dans la classe permet à certains de ceux-ci d'exprimer une facette différente de leur signification, assouplissant la contrainte sémantique exercée par les membres de la classe les uns sur les autres. Dans la mesure où, comme nous l'avons déjà précisé, c'est la force de cette contrainte interprétative croisée entre les mots d'une même classe qui nous garantit la qualité de la détection de thèmes réalisée par celle-ci, cet assouplissement n'est pas souhaitable ; d'où le choix effectué d'interrompre la croissance de la classe considérée *avant* l'ajout de ce nouveau terme.

Ce traitement est suivi d'un filtrage écartant les classes de mots-clés trop anecdotiques reconnaissant moins de 200 paragraphes, soit environ 2 ‰ du corpus — quelques classes, correspondant en général à un « sous-thème » extrêmement ciblé d'un thème déjà recensé par ailleurs, entrent dans cette catégorie —, filtrage au terme duquel on dispose d'environ quarante classes de mots-clés rassemblant trente à quarante mots chacune (la classe obtenue à partir du noyau « enseignement » donné en exemple à la section précédente est présentée en section 2.8).

Nommage des classes

Les classes de mots-clés obtenues sont toujours susceptibles, sous la forme définitive qu'elles ont dorénavant atteinte, d'une lecture complète permettant de comprendre le thème qui les sous-tend. Cette opération est néanmoins assez fastidieuse et il semble utile de disposer d'un nom synthétique les qualifiant de manière précise et directe — ce nom peut en outre par la suite être employé pour désigner les domaines associés, et à la limite les sèmes mésogénériques. Nous proposons donc une heuristique simple pour associer à chaque classe de mots-clés un triplet de mots significatifs désignant sans ambiguïté le thème qu'elle caractérise. Elle consiste à extraire de chaque classe de mots-clés $\mathcal{C}$ un sous-ensemble de trois mots tel que l'ensemble des paragraphes contenant au moins l'un de ces mots inclue une partie la plus étendue possible de

l'ensemble des paragraphes reconnus par $\mathcal{C}$. La sélection des mots est effectuée par recherche exhaustive sur toutes les combinaisons possibles de trois mots, approche rendue acceptable par le nombre relativement faible de mots à étudier.

La contrainte que nous avons définie est suffisante pour assurer que les trois mots choisis sont assez fréquents, et ne correspondent pas à une composante absolument anecdotique du thème caractérisé par $\mathcal{C}$. De plus, afin de maximiser la « couverture » des paragraphes reconnus par $\mathcal{C}$, chacun des trois mots sélectionnés est présent dans un nombre important de paragraphes où les deux autres n'apparaissent pas, et le triplet fournit donc un indice de « l'extension » du thème sous-jacent à $\mathcal{C}$. Ainsi une classe dont la dénomination extraite est < enseignement / école / université > reconnaîtra-t-elle les paragraphes évoquant le système éducatif (*enseignement*) au sens large, du primaire (*école*) au supérieur (*université*).

Nous procédons dans la section suivante à une présentation des résultats obtenus grâce au système FAESTOS, ainsi qu'à une évaluation de ceux-ci, en prenant en compte deux points de vue possibles sur la définition de la « qualité » d'une classe de mots-clés.

2.8. Présentation et évaluation des résultats

Les classes de mots-clés produites par le système FAESTOS peuvent être analysés selon deux critères : on peut d'une part s'interroger sur la valeur descriptive des classes de mots-clés et sur leur cohérence apparente pour un interprétant humain (évaluant ainsi le volet de notre système consacré à la caractérisation de thèmes) ; d'autre part, il est également important de contrôler leur efficacité en tant qu'outils de détection thématique. Nous étudions dans cette section les classes de mots-clés extraites par le système en adoptant successivement ces deux définitions possible de leur qualité.

2.8.1. Approche « intuitive »

Cette approche constitue la perception la plus directe de la qualité des classes de mots-clés obtenues, ainsi bien sûr que la plus subjective. Il nous a néanmoins semblé important, étant donné l'objectif fixé de caractérisation de thèmes, de contrôler que les résultats se prêtent volontiers à une appréhension humaine. De ce point de vue, les classes de mots-clés produites remplissent leur rôle et reflètent toutes clairement un thème précis, assez fidèlement exprimé par le triplet de mots faisant office de nom dont le mode de sélection est exposé à la section 2.7. Nous présentons ici le contenu de deux exemples de classes extraites par FAESTOS afin de permettre au lecteur de se faire une

idée du type de résultats produits, ainsi que la liste des « noms » des 39 autres classes générées au cours de cette même exécution[17] :

< **enseignement/école/université** > : { *alphabétisation, collège, diplôme, diplômé, école, éducatif, éducation, élève, enseigné, enseignement, enseignant, étudiant, étudié, instituteur, lycée, mathématique, primaire, pédagogique, scolaire, scolarisation, secondaire, universitaire, université* },

< **producteur/agriculture/céréale** > : { *agriculteur, alimentation, aliment, approvisionnement, autosuffisance, blé, bétail, coton, céréale, céréalier, denrée, élevage, engrais, famine, fruit, grain, huile, intensif, irrigation, lait, légume, nourriture, œuf, pain, pesticide, plantation, producteur, rendement, repas, riz, récolte, sac, saison, sec, semence, stock, sucre, surplus, sécheresse, vache, viande, vin, vivrier* },

< pape / foi / évêque >, < entreprise / investissement / infrastructure >, < film / cinéma / écran >, < unification / chancelier / est-allemand >, < pétrolier / gaz / baril >, < territoire / occupation / colonisation >, < roi / saoudite / chiite >, < durée / professionnel / qualification >, < écrivain / roman / revue >, < technologie / ingénieur / biologique >, < alliance / pacte / traité >, < congrès / constitution / suprême >, < crime / nazi / historien >, < informatique / satellite / machine >, < élection / voix / victoire >, < campagne / démocrate / candidat >, < théâtre / scène / musique >, < missile / conventionnel / arsenal >, < membre / réunion / comité >, < fascisme / rhétorique / négation >, < impôt / fiscal / assurance >, < télévision / image / journaliste >, < quartier / parc / tourisme >, < nation / conseil / international >, < armée / colonel / guérilla >, < déchet / énergétique / centrale >, < travail / emploi / salarié >, < réforme / ferme / propriétaire >, < règle / juridique / article >, < minorité / langue / ethnique >, < taux / baisse / croissance >, < paix / conférence / frontalier >, < financier / banque / crédit >, < asile / immigration / séjour >, < maladie / hôpital / médical >, < forêt / pollution / sécheresse >, < accord / barrière / douane >, < prison / judiciaire / détenu >, < déficit / dette / budgétaire >.

Un lecteur du *Monde diplomatique* pourra constater que les classes de mots-clés énumérées réalisent une couverture assez complète de la palette thématique développée par notre corpus. Il est particulièrement intéressant de noter les variations de granularité observables en fonction des domaines de prédilection de celui-ci : ainsi, si tous les « arts vivants » se trouvent rassemblés au sein de l'unique classe < théâtre / scène / musique >, on remarque un luxe de détails en ce qui concerne les questions économiques ou géopolitiques. Cette constatation montre à l'évidence l'intérêt d'une étude réalisée sur corpus sans présupposé sur la nature des résultats attendus.

[17] L'intégralité des classes de mots-clés obtenues au cours d'une exécution du système est présentée en annexe A.

Dans le même ordre d'idée, on peut observer l'émergence de thèmes liés à des faits particulièrement marquants de la période couverte par les archives étudiées, qui acquièrent de par le nombre de textes qui leur est consacré le statut de thèmes à part entière : ainsi la classe < territoire / occupation / colonisation > se réfère-t-elle au conflit israélo-palestinien, et < unification / chancelier / est-allemand > à l'unification de l'Allemagne. On peut également découvrir à la lecture des classes obtenues de nombreux indicateurs des choix « idéologiques » du *Monde Diplomatique,* telle la présence du mot *propagande* au sein de la classe < télévision / image / journaliste > ou celle de *lobby* dans < campagne / démocrate / candidat >, classe notamment centrée sur le processus électoral américain.

Les classes obtenues permettent donc, en identifiant la plupart des principaux thèmes du corpus et en jetant un premier éclairage sur leur mode de traitement, d'effectuer un « survol » du contenu du corpus, ce qui présente un premier intérêt en soi, par exemple comme aide à l'exploration de grandes quantités de données textuelles. Cette cohérence « intuitive » des classes de mots-clés n'offre néanmoins aucune garantie concernant leur efficacité pour la tâche de détection de thèmes que nous visons. Nous nous attachons donc maintenant à évaluer leur pertinence de ce second point de vue, plus aisément quantifiable.

2.8.2. Efficacité des classes de mots-clés pour la détection de thèmes

Cette section propose un ensemble d'indicateurs numériques permettant de juger la pertinence et la complétude de la détection de thèmes réalisée par nos classes de mots-clés. De manière classique, nous nous intéressons ici à la précision de la détection (proportion de détections de thèmes réalisées à bon escient) et à son rappel (proportion de détections automatiques par rapport à celles que l'on aurait souhaité observer). Le critère d'évaluation considéré est donc cette fois plus objectif, bien qu'il laisse encore une certaine marge d'interprétation à l'expérimentateur chargé de décider si le thème détecté dans un paragraphe y est réellement présent.

Outre le principe de cooccurrence de deux mots adopté jusqu'ici, nous avons été amené à développer d'autres critères de reconnaissance de thèmes afin d'améliorer la précision et le rappel de la détection. Nous les décrivons à la section suivante, avant de présenter les indices numériques et procédures expérimentales employés pour évaluer la qualité de la détection de thèmes (section 2.8.2.2), et d'exposer enfin les résultats chiffrés en section 2.8.2.3.

2.8.2.1. Critères de détection élaborés

Une variation élémentaire du premier critère de détection de thèmes par cooccurrence de deux mots-clés dans un paragraphe consiste à accroître la sélectivité de celui-ci en augmentant le nombre de mots-clés requis à trois. Si l'on

peut ainsi obtenir une fiabilité plus grande de la détection, il est probable que cela soit au prix d'une baisse importante du rappel de la détection.

Un troisième critère tentant de concilier précision et rappel a été mis au point en faisant usage d'un niveau de structuration de notre corpus journalistique non exploité jusqu'ici : son découpage en articles. En effet, si une classe de mots-clés reconnaît plusieurs paragraphes au sein d'un même article, il est vraisemblable que le thème caractérisé par cette classe relève de la problématique abordée par l'auteur dans ce texte, et la probabilité de le voir apparaître ailleurs dans l'article s'en trouve accrue. Inversement, il est assez peu courant de voir un thème apparaître de manière isolée dans un unique paragraphe d'un article — et si cela se produit, le vocabulaire spécifique à ce thème se trouvera concentré dans le paragraphe en question, augmentant ainsi le nombre de mots-clés potentiels. Pour refléter ce double principe, on souhaite rendre le critère de détection d'un thème donné plus sensible si ce thème apparaît déjà dans l'article étudié, et plus strict si ce n'est pas le cas. Un nouveau critère est donc défini de la manière suivante : partant d'une détection de thèmes par le critère de cooccurrence de trois mots-clés présenté ci-dessus, on étudie pour chaque article le nombre de paragraphes détectés comme abordant un thème donné. Si ce nombre est supérieur ou égal à deux[18], alors le seuil de détection pour ce thème dans l'article considéré est baissé à deux mots-clés, permettant potentiellement la détection du thème dans un plus grand nombre de paragraphes de l'article ; sinon, ce seuil est rehaussé à quatre et le critère devient donc plus strict, remettant en cause la détection du thème dans le paragraphe isolé.

Nous présentons maintenant la procédure de validation employée et les résultats observés en faisant usage de ces trois critères pour tenter de localiser dans le corpus les apparitions des thèmes caractérisés par les classes de mots-clés obtenues.

2.8.2.2. Procédure de validation

Trois indices numériques sont pris en compte afin de juger du succès de la détection de thèmes réalisée par les classes de mots-clés construites :

- la « couverture » du corpus, c'est-à-dire la proportion de ses paragraphes auxquels il a été possible d'attribuer au moins un thème. Cette valeur est également exprimée en termes de pourcentage de mots du corpus ;
- la précision de la reconnaissance de thèmes est calculée sur une sélection aléatoire de 1 000 paragraphes auxquels au moins un thème est attribué, en contrôlant manuellement pour chacun d'eux la présence du (des) thème(s) détecté(s), sans aucune exigence concernant l'importance

[18]Cette valeur est indépendante du nombre de paragraphes rassemblés dans l'article considéré, mais il serait sans doute profitable de la définir de manière plus fine en prenant en compte cette donnée.

du thème à l'échelle du paragraphe ;

– le rappel, enfin, n'a été calculé que pour le thème < enseignement / école / université > (désigné par « enseignement » dans la suite) en recherchant manuellement dans le corpus 100 paragraphes l'abordant en effet[19], et en contrôlant quelle proportion de ceux-ci est reconnue par la classe lui correspondant.

Nous avons nous-même réalisé le calcul de ces valeurs, qui se fonde donc plus sur une approche « de bon sens » du texte que sur une réelle expertise. Il est toutefois à noter que nous avons pu, par une lecture régulière du mensuel étudié, développer une assez bonne familiarité avec les thèmes développés dans celui-ci.

2.8.2.3. Indices qualitatifs numériques

Le tableau 2.1 rassemble les valeurs calculées pour chacun des critères de détection présentés en section 2.8.2.1 des indices de couverture, précision et rappel.

	Couverture		Précision	Rappel « enseignement »
	paragraphes	mots	(paragraphes)	(paragraphes)
Deux mots	66 %	70 %	55 %	65 %
Trois mots	32 %	40 %	85 %	35 %
Trois mots + structure	58 %	65 %	85 %	63 %

TAB. 2.1 – *Couverture, précision et rappel de la détection de thèmes réalisée, calculés pour les trois critères employés. Le critère désigné par « trois mots + structure » est celui faisant usage de la structuration en articles du corpus*

Une première constatation possible à la lecture de ce tableau est la différence assez importante existant entre les valeurs de la couverture du corpus exprimées en paragraphes et en nombre de mots. Cette divergence s'explique par le fait que plus un paragraphe contient de mots, meilleure est la prise qu'il donne à une détection de thème par cooccurrence de mots-clés. Ainsi, pour le troisième des critères présentés au paragraphe 2.8.2.1, on calcule que les paragraphes reconnus comme abordant au moins un thème sont en moyenne 35 % plus longs que les paragraphes sans thème détecté.

Le rappel mentionné, calculé pour un seul thème, n'a à l'évidence qu'une valeur indicative et est susceptible de varier d'un thème à l'autre. Il permet

[19]Cette sélection requiert la lecture d'environ 4000 paragraphes du corpus, ce qui explique que nous nous soyons restreint à une validation sur un unique thème, en n'employant que 100 paragraphes.

néanmoins de constater que le dernier critère décrit, qui prend en compte la structuration en articles du corpus, parvient comme souhaité à concilier précision et rappel de la détection : on n'observe aucune perte de précision par rapport au critère de cooccurrence de trois mots-clés, et une faible baisse du rappel par rapport au critère original ne faisant usage que de deux mots. C'est donc ce troisième critère qui est en définitive employé par le système FAESTOS pour réaliser la détection de thèmes, et c'est à celui-ci que nous ferons de manière systématique référence dans la suite.

2.8.3. Bilan

FAESTOS nous permet donc de générer de manière totalement automatique une caractérisation circonstanciée des principaux thèmes abordés dans notre corpus d'archives du *Monde diplomatique*, et de détecter avec un succès raisonnable les occurrences des thèmes ainsi recensés à l'échelle de paragraphes entiers. Les classes de mots-clés obtenues présentent un sens intuitif pour un utilisateur humain et, combinées avec les dénominations par triplet de mots que nous avons générées, remplissent donc leur rôle en ce qui concerne la caractérisation de thèmes, allant même au-delà : en fournissant des indices sur la manière dont sont abordés les thèmes qu'elles caractérisent, les classes de mots-clés donnent en effet un aperçu assez riche du contenu du corpus, du type de rhétorique qu'il développe, et même de ses choix « idéologiques ».

Du strict point de vue de la détection de thèmes, indépendamment du reste de notre travail, le système décrit ne rivalise toutefois ni en précision ni en rappel avec des travaux comme [FG02], qui réalisent une segmentation très fine du texte et localisent précisément les occurrences de thèmes. Cela est entre autres dû à notre choix de granularité de découpage du corpus par paragraphes entiers, et au mode de détection de thèmes employé, volontairement simple et fondé sur une approche lexicale de la langue. C'est d'ailleurs pourquoi nous ne présentons pas de comparaison de résultats avec d'autres travaux du domaine : les objectifs poursuivis présentent trop de différences pour qu'une telle comparaison puisse être significative. Néanmoins, les résultats obtenus sont particulièrement positifs si l'on considère la simplicité d'application du système, réellement automatique et ne requérant aucune donnée auxiliaire. Ils montrent le bénéfice possible pour certains types d'études sémantiques visant l'acquisition de connaissances synthétiques sur une collection de textes d'une approche statistique globale des données textuelles telle que celle que nous avons adoptée, qui fait notamment abstraction de l'aspect séquentiel du texte et des relations syntaxiques entre mots.

Au-delà de ces apports et de la place qu'il tient dans notre « chaîne de traitements » en trois étapes pour l'acquisition d'informations lexicales sémantiques, le système FAESTOS constitue par lui-même une solution potentiellement intéressante au problème couramment soulevé en TAL de la caractérisation et détection de thèmes. C'est pourquoi nous nous attachons également à

son développement en tant qu'outil indépendant et esquissons donc maintenant, si l'on peut dire « en aparté », un ensemble de pistes pouvant permettre d'accroître la précision et le rappel de la détection de thèmes réalisée, avant de revenir à la fin de ce chapitre à notre problématique originelle de structuration de lexique.

2.8.4. Évolutions envisageables du système FAESTOS

Des deux « facettes » du système — la construction de classes de mots-clés et leur usage pour la détection de thèmes —, c'est la deuxième qui appelle le plus d'améliorations, et en particulier le problème de rappel toujours assez faible de la détection. Nous avons déjà évoqué une évolution possible du critère de détection exploitant la structure en articles du corpus, qui consisterait à faire dépendre de la taille de chaque article étudié le nombre de paragraphes évoquant un thème donné au-delà duquel on considère ce thème récurrent comme relevant du cœur de la problématique abordée dans l'article. On peut dans le même ordre d'idées envisager d'étudier plus précisément la répartition dans l'article de ces paragraphes évoquant un même thème : sont-ils regroupés dans une partie de l'article ou dispersés sur toute sa longueur ? apparaissent-ils au début de celui-ci, à la fin ? Cette étude pourrait par exemple permettre la mise au jour d'un thème central de l'article, appelé à devenir le thème « par défaut » de ses paragraphes, ce qui augmenterait très sensiblement la couverture du corpus.

Une autre voie exploitable pour l'accroissement du rappel de la détection consisterait à faire un usage plus fin du graphe présenté à la figure 2.7 : celui-ci permet en effet d'identifier, outre les noyaux que nous en extrayons, des liens existant entre les thèmes correspondant à ceux-ci. Il est envisageable d'exploiter ces connexions de diverses manières, la plus élémentaire consistant à augmenter la probabilité d'apparition d'un thème si l'un de ses voisins dans le graphe est détecté à proximité, ou constitue le thème principal de l'article au sein duquel s'effectue la détection.

Il est également possible, même si cela dépasse le cadre de notre problématique initiale, de proposer des évolutions du système susceptibles de lui permettre de satisfaire des applications différentes, pouvant être considérées comme plus exigeantes : il serait ainsi envisageable de renoncer au postulat « d'atomicité » des paragraphes pour tenter de cibler plus précisément les lieux de manifestation de thèmes en leur sein. L'une des caractéristiques marquantes des résultats observés est en effet la proportion relativement élevée de paragraphes reconnus comme abordant plusieurs thèmes, qui représentent environ 36 % des paragraphes couverts : on compte en moyenne 1,5 thème par paragraphe, le maximum étant atteint par un long paragraphe détecté comme développant huit thèmes distincts (et les abordant en effet, comme un contrôle manuel a pu le confirmer). Bien que, dans le cadre de notre étude, cette « mul-

tithématicité » ne constitue aucunement un problème, on peut être tenté d'en savoir plus sur les paragraphes concernés. Ceux-ci se répartissent en trois catégories principales, certains pouvant appartenir à plusieurs d'entre elles :

- les paragraphes au sein desquels se succèdent de manière séquentielle plusieurs thèmes distincts, par exemple afin de réaliser une transition entre deux idées ou d'exprimer différents points de vue possibles sur une question ;
- ceux qui développent également des thématiques distinctes, mais de manière parallèle : par exemple, un texte traitant du budget de l'armée évoquera de manière répétée des questions budgétaires et militaires sans que l'on puisse réellement considérer qu'il parle de l'un *puis* de l'autre ou réalise un « aller-retour » d'un thème à l'autre (F. Rastier parle dans ce cas d'isotopies *entrelacées*) ;
- enfin, ceux qui abordent en réalité un unique thème composite et complexe, cas assez fréquent dans notre corpus : peut-on évoquer les émirats du golfe Persique sans évoquer conjointement monarchie, religion et pétrole ?

Nous ne disposons malheureusement pas, de par les choix inhérents à notre approche (ne pas faire usage de données textuelles intra-paragraphe), d'informations suffisamment fines pour décider à laquelle de ces trois catégories rattacher un paragraphe donné : cela requerrait en effet d'explorer la structure interne du paragraphe, dont nous avons choisi de faire abstraction. Il s'agirait néanmoins sans aucun doute d'une extension intéressante du système présenté ici, qui pourrait, le cas échéant, renvoyer à l'usage d'une méthode de segmentation « pure »[20] pour raffiner l'analyse des paragraphes relevant de la première catégorie.

Toutes ces considérations ne doivent pas nous faire perdre de vue l'objectif premier de ce système : réaliser une première structuration du lexique en classes mésogénériques, ou domaines, tout en découpant le corpus afin de regrouper les textes donnant lieu au développement d'un discours mettant en œuvre le vocabulaire relevant de ces domaines. Nous revenons maintenant à cette préoccupation initiale.

2.9. Retour à l'acquisition de lexique

Avant de présenter l'exploitation pratique des résultats obtenus pour notre tâche de construction automatique de lexiques sémantiques, nous introduisons ici une remarque concernant l'intérêt de ces résultats en tant qu'élément venant conforter le bien fondé de la théorie linguistique sur laquelle nous basons notre approche.

[20] C'est-à-dire sans processus de caractérisation thématique associé — on peut par exemple penser à des systèmes faisant exclusivement usage d'indices linguistiques signalant la présence d'une transition entre thèmes [LP95].

2.9.1. Intérêt linguistique

F. Rastier évoque dans [Ras96] l'une des principales critiques apportées à la théorie du mécanisme d'activation et inhibition de sèmes par présomption d'isotopie : l'isotopie se manifeste et se caractérise par la répétition d'un ou plusieurs sèmes dans un texte ; simultanément, le choix des sèmes qu'il est pertinent ou non de considérer dans l'interprétation d'une occurrence donnée d'un sémème dépend des isotopies auxquelles celui-ci est susceptible de prendre part. À moins de faire l'hypothèse d'une « intuition » d'isotopie, contradictoire avec le parti pris de description rationnelle des mécanismes interprétatifs adopté par F. Rastier, il semble que l'on se trouve ici face à la question de la poule et de l'œuf.

Nos résultats montrent qu'il est possible de faire apparaître des isotopies par simple analyse de surface des données textuelles, sans présupposé interprétatif, ce qui confirme la théorie d'une « auto-détermination » de la langue et du texte — autrement dit, la boucle de rétroaction apparemment infinie de l'interprétation lexicale sur l'interprétation textuelle trouve sa résolution au sein du système de la langue, et même du texte, sans qu'il soit nécessaire de faire appel à des déterminations pragmatiques ou transcendantales. Cela constitue un argument important en faveur de la conception selon laquelle la langue peut se déterminer sinon entièrement, du moins en grande partie, par sa manifestation observable.

Ainsi conforté dans nos présupposés, nous pouvons revenir à notre tâche originale de construction de lexique : le travail de détection de thèmes présenté dans ce chapitre a pour but de faire émerger une structuration en domaines de celui-ci. Nous présentons maintenant la manière dont nous passons du texte au lexique, du thème au domaine.

2.9.2. Structuration du lexique en domaines

Il est temps de procéder, comme nous l'annonçons dans l'introduction de ce chapitre, au retour dans le monde du lexique, et de « traduire » en termes de structures lexicales les phénomènes textuels mis au jour. La problématique est ici d'extraire du corpus le vocabulaire fortement associé à chacun des thèmes caractérisés, dont nous supposons qu'il constitue le domaine correspondant. Nous adoptons pour cette tâche le même critère que celui exploité à la section 2.7, mesurant la fréquence relative des mots étudiés dans les paragraphes évoquant un thème donné par rapport au corpus général. Soit m un lemme de mot apparaissant dans le corpus et T l'un des thèmes extraits de celui-ci. Nous définissons une mesure d'association a entre m et T par :

$$a(m,T) = \frac{\frac{n_T(m)}{nPar(T)}}{\frac{n_{total}(m)}{nPar(corpus)}} = \frac{n_T(m).nPar(corpus)}{n_{total}(m).nPar(T)} \tag{2.6}$$

où $n_T(m)$ est le nombre d'occurrences de m dans les paragraphes reconnus par la classe de mots-clés associée à T (c'est-à-dire détectés comme évoquant le thème T), $n_{total}(m)$ le nombre total d'occurrences de m, $nPar(T)$ le nombre de paragraphes reconnus comme évoquant T et $nPar(corpus)$ le nombre total de paragraphes dans le corpus étudié. $a(m, T)$ mesure donc le rapport entre la fréquence (par paragraphe) de m sur les paragraphes évoquant T et sa fréquence moyenne sur l'ensemble du corpus. Plus a est importante, plus m est spécifiquement employé dans le discours développé autour de T.

Contrairement à ce qui était le cas à la section 2.7, notre objectif n'est pas cette fois d'obtenir des mots discriminants, réellement spécifiques à T, mais seulement des mots liés à l'évocation de T de manière suffisamment forte pour que l'on puisse à bon escient les supposer porteurs du sème domanial associé. Nous retenons donc tous les mots apparaissant dans les paragraphes évoquant T deux fois plus fréquemment au moins que dans l'ensemble du corpus. Ce critère rassemble naturellement les mots-clés du thème, mais aussi de nombreux mots potentiellement polysémiques dont le sens se trouve fixé par le contexte, ainsi que des mots peu fréquents. Contrairement aux classes de mots-clés, les domaines peuvent avoir des mots en commun, et nous n'imposons donc pas de critère d'appartenance unique.

La figure 2.9 présente en guise d'exemple l'ensemble des noms répondant à cette contrainte pour le sous-corpus extrait lié aux « nouvelles technologies » (classe $<$ information / réseau / communication $>$) et apparaissant au moins dix fois dans celui-ci. Parmi les mots rassemblés, on peut notamment distinguer les mots fortement et sans ambiguïté liés au thème considéré (beaucoup de ceux-là font partie de la classe de mots-clés correspondante), les mots polysémiques désambiguïsés par le contexte thématique (*nœud, signal*), et ceux pour lesquels le sème domanial n'est qu'afférent et actualisé par l'usage (*bulle, détection, mot-clé* (*sic*)).

L'étape suivante de notre travail de construction de lexiques sémantiques consiste à définir au sein des domaines ainsi définis des classes sémantiques ou *taxèmes*. Suivant l'exemple de nombreux travaux menés sur cette problématique, nous étudions pour cela les contextes d'apparition des mots à traiter ; il est alors important que les contextes étudiés pour caractériser un mot m correspondent tous, ou du moins dans leur grande majorité, à des occurrences de m où il est employé dans un sens identique — en d'autres termes, à des occurrences de la forme textuelle m renvoyant à un même signifié (ou sémème), celui appartenant au domaine étudié. Nous découpons à cette fin le corpus d'étude en sous-corpus thématiques.

2.9.3. Découpage du corpus en sous-corpus thématiques

L'objectif de ce découpage est d'associer à chaque domaine un ensemble de textes où existe une forte présomption de voir les mots du domaine employés

information
réseau
communication
système
télévision
service
groupe
technologie
entreprise
image
satellite
programme
chaîne
informatique
machine
télécommunication
industrie
ordinateur
presse
média
recherche
outil
domaine
journal
jeu
téléphone
donnée
opérateur
espace
accès
ligne
diffusion
journaliste
autoroute
film
intelligence
firme
public
concurrence
type
agence
radio
qualité
utilisateur
publicité
publicitaire
vitesse
mode
compagnie
norme
logiciel
échelle
station
conception
câble
utilisation
équipement
infrastructure
écran
liaison
fusion
enjeu
cinéma
technique
transport
message
connaissance
monopole
émission
usage
planète
train
traitement
document
individu
lecteur
support
client
quotidien
géant
temps réel

service public
électronique
ère
audience
usager
déréglementation
transmission
dimension
courrier
matériel
innovation
téléspectateur
bataille
circulation
contenu
flux
distance
foyer
distribution
chercheur
toile
concept
mutation
industriel
site
éditeur
multimédia
concentration
maîtrise
univers
téléviseur
offre
fonctionnement
grâce
matière
laboratoire
canal
magazine
cassette
édition
lancement
dispositif
consommateur
fournisseur
parc
salle
minute
loisir
pentagone
renseignement
surveillance
antenne
mondialisation
performance
standard
direct
promotion
numéro
optique
concurrent
cyberespace
langage
transaction
marchandise
musique
manipulation
circuit
réalisation
ingénieur
reportage
lanceur
sport
réglementation
caméra
fichier
lecture
version
atelier
programmation
simulation
vidéo

marketing
annonceur
électricité
Japonais
vol
taille
futur
actualité
utopie
laser
multiplication
fabrication
vice-président
permanence
secret
tarif
commande
fiction
réel
observation
biotechnologie
matériau
automobile
bibliothèque
fabricant
constructeur
rapidité
orbite
micro-ordinateur
photographie
disque
synthèse
robot
divertissement
rentabilité
paysage
filière
relais
métier
seconde
censure
anglais
photo
savoir
complexité
connexion
actionnaire
gamme
axe
rédacteur
transparence
chemin de fer
journalisme
nippon
accélération
licence
haute définition
minitel
clientèle
invention
rachat
studio
magnétoscope
robotique
messagerie
savoir-faire
navette
central
terminal
téléphonie
professionnel
acquisition
filiale
précision
commercialisation
réception
présentateur
serveur
pointe
routier
enregistrement

informatisation
multinationale
exposition
rédaction
engin
apprentissage
spot
navigation
synergie
internaute
frais
partenariat
comté
valeur ajoutée
vie privée
signal
flot
cerveau
démonstration
domicile
procédé
génie
conglomérat
globalisation
architecture
étage
feuilleton
ingénierie
console
japonais
directeur général
boîte
archive
vedette
onde
bit
interactivité
coin
pari
avènement
finalité
francs
bulletin
globe
fréquence
interconnexion
piste
humain
rationalité
pénétration
expérimentation
excellence
fibre
distributeur
imagerie
abonnement
capteur
numérisation
écologie
présentation
particulier
commercial
adresse
consortium
créativité
outre-Atlantique
recherche-développement
maintenance
retransmission
concepteur
débit
maillage
fibre optique
généralisation
profil
stockage
désinformation
séquence
interaction
corporation

star
correspondance
segment
récepteur
cybernétique
fax
télécopieur
zine
télétravail
plate-forme
diversification
entrave
reproduction
logistique
irruption
chimie
décor
pionnier
chaussure
densité
panne
machine-outil
détection
puce
interface
cd-rom
câblo-opérateur
grandeur
stratège
abondance
interdépendance
pilote
métro
joueur
alerte
disponibilité
petit écran
automatisation
informaticien
continu
mastodonte
microprocesseur
clavier
automate
fusée
discrétion
attraction
dessin
générateur
créneau
arbitre
catalogue
relations publiques
photographe
express
régie
réfrigérateur
annuaire
inforoute
processeur
créateur
opportunité
salon
consolidation
match
nouveauté
fascination
rayon
grille
collecte
rubrique
management
imprimerie
intimité
rai
télédétection
cédérom
modem
télévirtualité
boutique
rendez-vous

amont
originalité
tirage
francophonie
prolongement
carrefour
major
sou
configuration
majeur
fabrique
biologie
casino
animation
copie
lune
compression
conditionnement
noeud
ramification
pirate
gadget
science-fiction
tournage
dépêche
bouquet
météo
télex
hypertexte
télégraphe
géographie
investigation
valorisation
hexagone
panoplie
espionnage
biosphère
label
documentation
saut
exclusivité
micro
fiabilité
show
affichage
déontologie
vapeur
compatibilité
transfrontière
médium
vidéodisque
scientifique
spécialisation
expertise
bulle
paquet
ressource humaine
revient
comptabilité
paramètre
rayonnement
rareté
index
assemblage
éclairage
résumé
dérégulation
clip
interception
mot-clé
loterie
téléfilm
pilotage
mandarin
confidentialité
araignée
décodeur
copyright
cybermonde

FIG. 2.9 – *Liste des noms liés au domaine des « nouvelles technologies » apparaissant plus de dix fois dans le sous-corpus correspondant, présentée par ordre de nombres d'occurrences décroissants*

dans le sens qui est le leur, dans le lexique, au sein de ce domaine. Cette collection de textes est naturellement le sous-corpus rassemblant les paragraphes évoquant le thème « dual » du domaine, dont la construction est immédiate étant donné l'outil que nous avons mis en place.

Les sous-corpus extraits de cette façon sont de tailles très variables suivant l'importance de l'évocation de leur thème dans le corpus, ainsi que l'efficacité de la classe de mots-clés associée pour réaliser leur détection. On observe ainsi des tailles allant de quelques dizaines à quelques centaines de milliers de mots. Comme nous le voyons au chapitre suivant, la quantité de données disponible pour réaliser l'analyse des modes d'emploi des mots est un facteur conditionnant largement la qualité des résultats obtenus lors de la tâche de construction de taxèmes. C'est pourquoi nous sommes amené à limiter notre étude aux quelques domaines bénéficiant d'un sous-corpus associé de plusieurs centaines de milliers de mots.

2.10. Conclusion

L'évaluation détaillée du système FAESTOS ainsi que l'étude de ses extensions possibles ayant été abordées en profondeur à la section 2.8, nous nous limitons dans cette conclusion à quelques rapides remarques sur les apports de ce travail.

Les méthodes d'analyse des données textuelles que nous avons présentées dans ce chapitre constituent une solution satisfaisante au double problème de structuration du lexique en classes sémantiques mésogénériques (domaines) et d'association à celles-ci d'ensembles de textes particulièrement représentatifs. Les domaines construits sont en outre « étiquetés », ou nommés, ce qui constitue un surcroît d'expressivité appréciable pour la construction de lexiques sémantiques. Les recherches menées ont été l'occasion de la mise au point d'algorithmes originaux, proposant notamment une réponse à la question délicate de l'extraction de classes d'un arbre de classification sans *a priori* concernant le nombre ou la taille de celles-ci. Ces techniques nous permettent de construire un système totalement automatique ne reposant sur aucun présupposé concernant le corpus étudié, et donc potentiellement tout à fait portable.

Ces premiers résultats nous permettent d'aborder sur des bases solides la deuxième tâche à accomplir dans le traitement de notre problématique d'acquisition automatique sur corpus d'informations lexicales fondées sur la sémantique différentielle : le découpage des domaines en classes sémantiques de mots de sens proches, les *taxèmes* de la sémantique différentielle. Le chapitre suivant est consacré à la présentation des techniques que nous avons mises au point afin de proposer une solution à cette question.

Chapitre 3

Construction de classes sémantiques

Prodicos me reprend toujours quand pour te louer, toi ou un autre, je dis « Protagoras est un savant et terrible homme », et me demande si je n'ai pas honte d'appeler terribles les choses qui sont bonnes. Terrible désigne selon lui quelque chose de mauvais : en effet, on ne dit jamais « terrible richesse », « terrible paix », « terrible santé », mais on dit « terrible maladie », « terrible guerre », « terrible pauvreté » ; ainsi, ce qui est terrible est mauvais.

Platon, *Protagoras* [Pla67]

3.1. Introduction

Selon la hiérarchie de représentation du sens développée par la sémantique différentielle, les domaines, qui constituent le premier niveau de structuration du vocabulaire, sont à leur tour partitionnés en taxèmes. Ceux-ci reflètent une organisation paradigmatique du lexique contrainte par l'axe syntagmatique : en contexte, cette contrainte syntagmatique définit l'ensemble des mots « que l'on aurait pu employer à la place de X » à une position donnée dans un énoncé ; dans le lexique, un taxème est une telle alternative fonctionnellement et socialement normée. Un taxème ne rassemble donc pas nécessairement que des synonymes ou quasi synonymes, mais aussi potentiellement des antonymes, hyponymes ou hyperonymes. Nous présentons dans ce chapitre le travail accompli afin de découper les domaines générés au chapitre précédent en classes sémantiques correspondant, dans l'idéal, à ces taxèmes, toujours à partir des seules données textuelles rassemblées dans notre corpus d'étude.

Le principe employé pour la construction automatique de classes sémantiques, dans le cadre de nos travaux comme dans de nombreuses recherches antérieures, découle directement de la définition contextuelle que nous venons d'en donner : il consiste à rassembler des mots employés de manière analogue

dans les données étudiées, en faisant l'hypothèse que cette équivalence fonctionnelle entre mots implique une proximité réellement sémantique [MC91]. L'une des principales inspirations de cette approche peut être trouvée dans les travaux de Z. Harris, qui défend une approche formaliste, « mathématique » de la langue [Har68, HGR+89], fondée uniquement sur l'étude d'énoncés élémentaires dans un domaine précis, indépendamment de leur contexte de production, de la réalité à laquelle ils font référence, ou de leur effet sur le locuteur et le destinataire :

> « *The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.* »[1] (Z. Harris, [Har68])

Dans [Har51], il illustre son approche par l'exemple d'une langue fictive entièrement définie par six énoncés :

1. La lune brille.
2. Le lac brille.
3. La lune brille dans l'eau.
4. La lune est rouge.
5. Le lac est sale.
6. La lune paraît sale.

D'après (1) et (2), « la lune » et « le lac » appartiennent à une même classe d'objets linguistiques, puisqu'ils peuvent être substitués l'un à l'autre sans modification du contexte « brille ». De (1) et (3), on déduit que « dans l'eau » constitue un complément facultatif. D'après (4) et (5), et en vertu de l'équivalence démontrée de « la lune » avec « le lac », « rouge » et « sale » appartiennent à une même classe. De même, les énoncés (5) et (6) nous permettent enfin de réunir « est » et « paraît ». Ce type d'analyse a encore cours notamment dans le domaine de l'inférence grammaticale pour le traitement de sous-ensembles restreints et très codifiés de la langue.

En pratique, néanmoins, ce test de permutabilité, qui n'est valide que pour des énoncés identiques (ou du moins fonctionnellement identiques), représente une contrainte beaucoup trop forte pour être exploitable tel quel : les énoncés potentiels de la langue sont, sinon innombrables, du moins indéfiniment nombreux, et aucun corpus ne peut prétendre en recenser ne serait-ce qu'une part conséquente. Il est donc nécessaire d'amender ce principe d'identité fonctionnelle des contextes d'usage des mots pour adopter celui de « ressemblance », plus souple mais également plus délicat à quantifier. Pour chacun des mots étudiés, la plupart des travaux de génération de classes sémantiques

[1] « La signification des entités [linguistiques, NdT], comme la signification des relations grammaticales qu'elles entretiennent, est liée à la restriction des combinaisons de ces entités vis-à-vis d'autres entités. »

menés à bien construisent donc une caractérisation synthétique de son usage tel que constaté par l'observation de ses différentes occurrences dans le texte. Ils définissent ensuite une similarité entre mots faisant usage de cette caractérisation de leurs emplois dans le texte, laquelle est exploitée pour regrouper les mots en classes. Il est remarquable que ce passage de l'exigence d'identité à celle de ressemblance nous rapproche des principes de la sémantique interprétative : dans cette théorie, en effet, s'il est possible d'employer un mot à la place d'un autre dans un énoncé donné, c'est parce que le mot substitué est susceptible d'entretenir le même type de rapport sémantique vis-à-vis des mots environnants que l'original ; en particulier, il est capable de participer à divers degrés de localité aux mêmes isotopies que celui-ci. En considérant la similarité entre les contextes d'usages de deux mots, l'important est donc moins la « surface » de ces contextes que les contraintes sémantiques qu'ils imposent, définissant ainsi « la forme de la case » dans laquelle le mot considéré doit s'insérer.

Comme nous l'avons évoqué en introduction de ce mémoire, nombreux sont ceux parmi les travaux existants dans le domaine de l'acquisition d'informations lexicales, et notamment de classes sémantiques, qui font usage de corpus techniques : ceux-ci relèvent d'une langue où l'élégance de l'expression est secondaire par rapport à la précision et à la concision du texte, et dans laquelle il existe le plus souvent une manière consacrée d'exprimer chacune des idées-clés du domaine. Nous travaillons pour notre part sur un corpus généraliste dont la langue est particulièrement riche (voire inventive) et le style travaillé, évitant notamment les répétitions et faisant usage de toutes les ressources de vocabulaire à disposition des auteurs. Il s'agit là bien entendu d'une source de difficulté importante, puisqu'il est plus délicat de rapprocher les mots utilisés de façon similaire lorsque les auteurs font de leur mieux pour éviter d'employer de manière répétitive des formulations analogues. En revanche, nous pouvons grâce au découpage du corpus du *Monde diplomatique* en sous-corpus thématiques présenté au chapitre précédent profiter, au moins partiellement, de l'un des avantages inhérents à l'usage de corpus spécialisés, à savoir la stabilité des sens des mots employés au sein de chacun de ces sous-corpus. De plus, conformément aux principes théoriques de la sémantique différentielle, nous ne réalisons de structuration en classes sémantiques qu'au sein des domaines constitués à l'étape précédente, ce qui garantit en théorie que les textes considérés sont pertinents pour caractériser les mots étudiés.

Contrairement à la question de détection de thèmes traitée au chapitre précédent, à laquelle il est possible de proposer une réponse totalement automatisée satisfaisante et « définitive », le problème de la construction automatique à partir de corpus de classes sémantiques ne connaît pas à l'heure actuelle de telle solution. Les recherches existant dans ce domaine, dont nous décrivons les plus intéressantes et représentatives à la section 3.2, permettent en effet toutes de faire apparaître des rapprochements entre mots sémantiquement pertinents, mais s'en remettent toujours à une évaluation humaine ou à

un thesaurus préexistant pour interpréter et « filtrer » les résultats obtenus. Le système que nous présentons dans ce chapitre n'échappe pas à cette règle, mais n'en propose pas moins plusieurs avancées notables par rapport aux travaux antérieurs. D'une part, conformément aux impératifs généraux que nous nous sommes fixés, il ne fait usage que de techniques d'analyse de données non supervisées exploitant exclusivement le corpus de texte brut étiqueté, sans analyse syntaxique — prérequis de beaucoup de recherches menées jusqu'ici — ni exploitation de données auxiliaires. D'autre part, les techniques que nous avons développées permettent de constituer des classes sémantiques à partir de corpus de langue non spécialisée de relativement petite taille. En effet, le découpage du corpus d'étude original en sous-corpus thématiques, qui nous permet de n'imposer aucune contrainte de type de langue concernant les textes étudiés et donc d'assurer la généricité de notre travail dans son ensemble, a pour conséquence que les sous-corpus à partir desquels doivent être constituées les classes sémantiques sont de volume très restreint : les plus importants d'entre eux rassemblent quelques centaines de milliers de mots, ce qui est une taille courante pour l'analyse de langues de spécialité mais pas pour des corpus de « langue générale », fussent-ils thématiquement cohérents. La problématique déjà évoquée de faible représentativité des données textuelles par rapport à la complexité de la langue est donc un facteur particulièrement limitant dans le cadre de notre travail, et nous sommes amené pour la surmonter à développer des méthodes d'analyse de données textuelles originales permettant d'optimiser la « rentabilité » de l'exploitation du peu de textes dont nous disposons pour comparer les emplois de mots — tout en restant, toujours selon un principe de généricité maximale, applicables à des corpus de tailles plus importantes.

Nous précisons à la section suivante le positionnement de nos recherches vis-à-vis de la grande famille des travaux de constitution sur corpus de classes sémantiques. La section 3.3 est ensuite consacrée à un approfondissement de la problématique considérée : nous y présentons les enseignements tirés des nombreuses expériences réalisées avant d'aboutir au système finalement proposé, ainsi que les principes généraux de ce dernier. Sa description détaillée fait l'objet des sections 3.4 et 3.5, et est suivie en conclusion (section 3.6) d'un retour aux principes linguistiques de notre étude.

3.2. Travaux existants

La question de la construction automatique de classes sémantiques par analyse de corpus a donné lieu à de nombreuses recherches explorant des méthodes très variées. Nous en donnons dans ce qui suit un aperçu rapide, en nous restreignant volontairement au domaine des travaux à portée purement lexicale — en particulier, nous ne faisons qu'exceptionnellement mention de la littérature tout aussi fournie concernant l'acquisition automatique d'onto-

logies, en accord avec la distinction sur laquelle nous insistons au cours de la section 1.2.3.

La formalisation proposée par G. Grefenstette [Gre93, Gre94] des affinités du premier, deuxième et troisième ordre entre mots, déjà évoquée au cours de l'introduction à ce document, nous sert ici de « guide de lecture ». Les affinités du premier ordre rapprochent les mots apparaissant de manière conjointe « anormalement » fréquemment. Rassembler des mots selon un critère d'affinité du premier ordre (par exemple en mesurant l'information mutuelle* existant entre eux) permet par exemple de mettre au jour des termes complexes ou expressions figées ; il s'agit d'un type de technique largement exploité en recherche de terminologie. Les affinités du deuxième ordre rassemblent pour leur part les mots employés dans des contextes similaires. On retrouve ici l'idée fondamentale déjà exposée, issue des principes de la linguistique harrissienne ; G. Grefenstette précise néanmoins à la suite de Harris [HGR$^+$89] que, hors du cadre très balisé d'une langue de spécialité, cette proximité fonctionnelle entre mots n'équivaut pas à une similarité sémantique, à cause des problèmes soulevés par la possible polysémie des mots étudiés. Les affinités du troisième ordre, enfin, sont introduites afin de prendre en considération et résoudre ce problème d'ambiguïté des objets traités. Plutôt que de comparer des mots, G. Grefenstette s'intéresse à la notion d'axe sémantique, définie comme une paire de mots réciproquement proches (au sens des affinités du deuxième ordre) l'un de l'autre. L'ensemble des mots apparaissant à proximité des deux mots déterminant un tel axe compose une classe sémantique à proprement parler, l'ambiguïté sur les sens des mots étant levée par la condition de double similarité avec les « extrémités » de l'axe.

Les affinités du deuxième ordre constituent le fondement de la majorité des travaux de construction de classes sémantiques existants — peu d'auteurs ont en effet exploité le principe des affinités du troisième ordre, faisant le choix soit d'exploiter des corpus spécialisés, soit d'ignorer la limitation théorique de validité des méthodes du deuxième ordre. Nous présentons à la section 3.2.1 cette première famille de recherches, à laquelle se rattachent également les nôtres. Il ne s'agit pas de la seule approche possible du regroupement sémantique de mots : certains travaux font, à la suite de [Hea92], le choix de construire les classes sémantiques par extraction de motifs linguistiques particuliers de cooccurrence de mots — ce qui se rapproche par la forme d'affinités du premier ordre, l'objectif étant toutefois la mise au jour de relations paradigmatiques. Nous présentons les études relevant de cette seconde catégorie à la section 3.2.2. Pour finir, la section 3.2.3 introduit à la lumière de ces diverses références les choix fondamentaux ayant dirigé nos recherches.

3.2.1. Affinités du deuxième ordre

Les travaux présentés ici se placent tous plus ou moins directement dans la lignée des principes linguistiques de Z. Harris, et leur méthodologie s'inspire

de celle exposée par D. Hindle [Hin90] et F. Pereira *et al.* [PTL93], qui s'intéressent à la classification de noms en caractérisant chacun d'eux par l'ensemble des verbes dont il est le plus couramment un objet direct. R. Agarwal choisit dans [Aga95] d'ajouter à cette information d'usage la connaissance des verbes dont le nom est le plus souvent le sujet, ainsi que des adjectifs le plus fréquemment employés pour préciser son sens. La classification hiérarchique des noms réalisée grâce à ces données est soumise à validation par plusieurs experts du domaine afin de constituer des classes sémantiques à proprement parler. Ce travail est assez représentatif de nombreuses recherches du domaine, présentant des caractères similaires : il fait usage de corpus spécialisés, exploite des liens syntaxiques pour caractériser les contextes d'usage, et adopte sur la question du sens une perspective aussi « taxonométrique » que lexicale, dans la mesure où le jugement final est du ressort de l'expert du domaine, et non du linguiste.

Donnant toujours une place prépondérante aux verbes pour la caractérisation des noms, le système ASIUM développé par D. Faure et C. Nédellec [FN98] analyse des représentations complètes d'énoncés verbaux syntaxiquement structurés : à la différence des travaux précédents, les compléments indirects sont également retenus, et typés par la préposition qui les introduit. Le système collecte ainsi sur l'ensemble du corpus des instanciations de structures argumentales de verbes, et procède au regroupement de noms en classes permettant une généralisation optimale de schémas de ces structures. ASIUM obtient des résultats de bonne qualité, mais ne fonctionne que sur des textes relativement « simples » (recettes de cuisine, documents techniques, *etc.*), ce qui s'explique en particulier par sa forte dépendance vis-à-vis de la qualité de l'analyse syntaxique réalisée.

D. Lin et P. Pantel s'intéressent pour leur part à l'étude de corpus de langue plus « ordinaire » (journalistique, en l'occurrence) et présentent dans [LP01] l'algorithme UNICON (*UNsupervised Induction of CONcepts*), qui construit un ensemble de *clusters* de noms selon une mesure de similarité fondée sur la comparaison des relations syntaxiques entretenues par les noms considérés avec les autres mots du corpus [Lin98]. L'algorithme décrit est conçu afin de pouvoir traiter des données « imposantes », tant en terme de volume que de complexité dimensionnelle de leur représentation. Il permet d'atteindre par étude de corpus d'au moins 50 millions de mots des taux de classification pertinente variant de 70 à 85 %.

Le système ZELLIG, initialement présenté par B. Habert, E. Naulleau et A. Nazarenko dans [HNN96], permet l'étude de relativement petits corpus spécialisés (quelques dizaines ou centaines de milliers de mots). Ses auteurs font usage des termes complexes extraits du corpus par l'extracteur de terminologie LEXTER [Bou94], dont ils simplifient la structure syntaxique afin d'en extraire des arbres de dépendance élémentaires. Ils construisent à partir de ces données un graphe dont les sommets sont les mots étudiés, et dont

les arcs sont étiquetés par les contextes élémentaires qu'ils partagent, l'intuition largement confirmée par l'expérience étant que les arcs portant un grand nombre de contextes communs relient des mots de sens proches. L'exploitation de ce graphe reste principalement manuelle : l'extraction de classes est réalisée par recherche de cliques* ou de composantes connexes, après seuillage par un nombre de contextes communs minimal défini arbitrairement. Les auteurs s'intéressent principalement dans la suite de leurs expériences à l'utilisation des résultats pour la construction d'ontologies : [BHNZ97, BHNZ00, BHNZ01] explorent les possibilités d'usage des techniques mises au point pour l'amendement d'une terminologie médicale « généraliste » à partir de textes médicaux spécialisés. L'aspect plus proprement lexical de l'exploitation possible du système ZELLIG est notamment présenté dans [FHL97], étude s'intéressant au graphe de proximités sémantiques obtenu par analyse d'un corpus textuel rassemblant les interventions radiotélévisées prononcées par F. Mitterrand au cours de son premier mandat présidentiel (ce corpus rassemble environ 300 000 occurrences de mots). L'expression étant beaucoup moins rigide que dans le corpus médical précédemment mentionné, et les mots employés beaucoup plus susceptibles de polysémie, le graphe obtenu rapproche cette fois des mots de sens *liés*, mais pas nécessairement réellement proches, comme *pays* et *gouvernement* ou *politique* et *affaire*. L'un des apports majeurs de cette étude est de montrer que le graphe construit par ZELLIG permet d'identifier différentes « facettes » sémantiques des mots qu'il rassemble, ce qui suggère la possibilité d'une extension des capacités du système au-delà de la simple construction de classes sémantiques ; cette identification suppose toutefois un travail d'interprétation humaine assez important.

La construction de classes sémantiques est le plus souvent entreprise pour étudier les noms, mais certains travaux se focalisent également sur la classification des verbes. L'objectif n'est plus cette fois la construction de classes sémantiques *stricto sensu*, mais le rassemblement de verbes présentant des structures argumentales similaires (présence ou absence de complément d'objet direct, indirect, d'indication directionnelle, *etc.*) : B. Levin suggère dans [Lev93] qu'une classification des verbes en fonction de ce critère implique une catégorisation sémantique. Cette hypothèse est mise à l'épreuve de l'acquisition automatique (mais supervisée) dans [DJ95], B. Dorr et D. Jones montrant également par la suite que la catégorisation obtenue est compatible avec celle de WordNet [DJ96]. Ces premières études ont donné lieu à de nombreuses suites, comme [OM97], qui montre la pertinence de l'approche pour une langue non occidentale (le japonais), ou [MS01], qui explore les possibilités d'apprentissage non supervisé. Remarquons que l'objectif et la méthodologie de ces travaux sont les « réciproques » de ceux adoptés pour le système ASIUM.

Les recherches citées jusqu'à présent dans cette section font toutes usage d'une analyse syntaxique plus ou moins poussée des textes étudiés, ce qui constitue le cas de figure le plus courant. Certains travaux évitent néanmoins cette contrainte, en particulier ceux réalisant la construction de classes devant

permettre une modélisation « n-classe » [BPd$^+$92] du langage : l'objectif est alors de construire des classes de mots telles que la connaissance des probabilités de transitions entre elles permette une approximation optimale des probabilités de transitions entre les mots qu'elles regroupent. Travaillant dans le domaine de la reconnaissance d'écriture, J. Hughes et E. Atwell proposent dans [HA94] de rassembler les mots d'un corpus de texte en comparant les listes de ceux apparaissant dans les voisinages de deux mots précédant et suivant leurs occurrences. L'objectif original de ce travail est la mise au jour de catégories morphosyntaxiques, mais malgré l'aspect « rudimentaire » des méthodes employées (corpus de texte brut, mesures statistiques simples), des regroupements sémantiques de mots sont clairement observables dans la classification produite. Des résultats similaires sont observés dans [MS96] par F. Smith et J. McMahon, qui poursuivent des recherches analogues dans le domaine de la reconnaissance vocale.

Un autre domaine ayant donné lieu à la génération « incidente » de classes sémantiques est celui de l'induction de grammaires formelles de la langue. Dans [SM99], K. Siu et H. Meng génèrent à partir de phrases-exemples des règles de réécriture du type de celles employées dans les grammaires hors contexte, en « entrelaçant » des rapprochements de mots selon des affinités premier et deuxième ordre. Sont ainsi obtenues des règles de réécriture telles que :

$$\begin{array}{lcl} SC1 & \rightarrow & \text{monday} \mid \text{tuesday} \\ SC2 & \rightarrow & \text{morning} \mid \text{afternoon} \mid \text{evening} \\ TC1 & \rightarrow & SC1\ SC2 \\ SC3 & \rightarrow & \text{now} \mid \text{today} \mid TC1 \end{array}$$

Le non-terminal $SC3$ génère une série d'expressions comme « monday morning », « tuesday afternoon », « now », *etc.*, que nous pourrions désigner comme une « classe sémantique d'expressions ». Le système présenté, en combinant deux types d'affinités, permet ainsi de constituer des classes sémantiques beaucoup plus riches que les approches purement paradigmatiques. Malheureusement, les expériences que nous avons pour notre part réalisées afin de tenter d'appliquer une technique similaire à notre problématique tendent à montrer que ce type de méthode ne peut fonctionner que pour des langues très simples et spécialisées, disposant d'un vocabulaire limité.

La section suivante est consacrée à la description de travaux poursuivant le même objectif que ceux déjà présentés, mais se distanciant du principe des affinités du deuxième ordre pour faire usage de méthodes d'analyse des données textuelles plus symboliques.

3.2.2. Recherche de motifs linguistiques

Tous les travaux évoqués dans cette section tirent profit de certaines figures de langage qui conduisent l'auteur d'un texte à rapprocher explicitement

dans celui-ci des mots appartenant aux mêmes classes sémantiques. L'une des premières à mettre en œuvre une telle méthodologie est M. A. Hearst, qui propose dans [Hea92] de détecter sur corpus des paires de mots en relation d'hyponymie / hyperonymie (l'objectif n'est donc pas cette fois la construction de classes sémantiques), en s'inspirant de techniques employées pour l'exploitation de définitions issues de dictionnaires informatisés [MAE86, NN88]. Celles-ci consistent à exprimer sous forme d'expressions rationnelles des patrons d'extraction qui capturent des formulations rapprochant typiquement des mots liés par la relation sémantique recherchée. En français, pour la relation d'hyponymie, des patrons tels que

$$N,\ \text{une sorte de}\ N$$

ou

$$\text{des}\ N\ \text{tels que}\ (N,\ )*\ N\ (\text{et}\,|\,\text{ou})\ N$$

seront ainsi définis. Les expériences menées sur un corpus journalistique de 20 millions de mots montrent que les résultats obtenus par cette méthode sont de bonne qualité, mais souffrent d'un faible « rendement » dans l'exploitation des données disponibles : les règles définies ne permettent en effet d'extraire que quelques dizaines de paires de mots en relation d'hyponymie / hyperonymie. Des études récentes plus poussées menées dans cette même direction [MZB04] suggèrent que s'il est possible d'augmenter ce « rendement » en variant les motifs employés, cela ne peut se faire qu'au prix d'une perte en précision.

Les recherches menées par E. Riloff et J. Shepherd [RS97, RS99] adoptent la même philosophie que ce premier travail, mais dans un but différent : l'objectif n'est pas de rechercher une relation sémantique particulière, mais de construire des classes sémantiques de mots de sens proches — entre autres hyponymes et hyperonymes. Les motifs d'extraction employés reflétent des phénomènes linguistiques plus simples que ceux exploités dans [Hea92] : énumérations (listes et conjonctions), formules appositives (« *the horse, a black stallion* ») et mots composés (« *tuna fish* », « *oak tree* »). La méthode présentée permet la construction de classes de taille importante (plusieurs dizaines de mots) à partir de « graines » (prototypes de classes rassemblant quelques mots), mais requiert un important travail de filtrage manuel. De plus, les auteurs reconnaissent que la méthode décrite n'est probablement adaptée qu'à l'extension de classes fortement représentées dans le corpus étudié. Dans [CR98], E. Charniak et B. Roark proposent une évolution de [RS97] employant des indices statistiques plus élaborés, et obtiennent ainsi des résultats significativement supérieurs.

Le système NOMEN, présenté dans [LYG02, LYG03] a également pour but d'étendre des classes dont quelques exemples initiaux (« graines ») sont fournis manuellement. Il se rapproche dans son principe des travaux décrits précédemment, mais réalise automatiquement sur corpus l'apprentissage des motifs discriminants permettant d'acquérir de nouveaux éléments d'une classe.

Ceux-ci se trouvent ensuite étendus et raffinés par un processus de *feedback* à mesure que la classe grandit. Le système n'est malheureusement évalué que pour l'acquisition de classes de noms propres, catégorie pour laquelle la tâche d'acquisition de classes sémantiques est grandement facilitée par les formulations du type « Madame X, directrice de… », « Monsieur Y, ouvrier spécialisé », très fréquentes en particulier dans les corpus journalistiques ; il serait en particulier intéressant de connaître le bénéfice du système présenté par rapport à une approche simple utilisant ces quelques patrons élémentaires.

Citons encore dans cette catégorie les recherches menées par D. Kokkinakis *et al.* [KGW01], qui étendent et spécialisent un lexique sémantique du suédois structuré par des classes sémantiques en étudiant les associations de mots de ces classes avec des mots non classés au sein d'énumérations. La méthode employée est ainsi très similaire à celle de [RS97], profitant en outre de plus de connaissances sémantiques définies *a priori* (l'ensemble des classes sémantiques du lexique étudié), et les résultats obtenus par les auteurs sont, sans surprise, d'excellente qualité.

Le principal défaut pratique des travaux évoqués dans cette section est une très forte dépendance au volume de données disponible pour réaliser l'apprentissage. Cette limitation est en particulier soulignée dans [DW02], qui présente une recherche similaire à [RS97] et [CR98]. B. Dorow et D. Widdows y mettent en parallèle la qualité de leurs résultats avec le facteur de taille du corpus exploité : le travail présenté dans [RS97] obtient une précision de 17 % en employant un corpus de 500 000 mots, celui de [CR98] une précision de 35 % sur 30 millions de mots, et enfin celui de [DW02] une précision de 82 % sur le *British National Corpus*, qui compte 100 millions de mots. Dans l'optique de notre recherche, où les données textuelles disponibles sont restreintes à quelques centaines de milliers de mots, de telles exigences de volume rendent à l'évidence « impraticables » les techniques de cette famille.

Nous retenons également comme réticence vis-à-vis de ces méthodes l'aspect arbitraire et « artificiel » de l'exploitation de formules linguistiques particulières rendant manifeste sur l'axe syntagmatique du texte sa dimension de sélection paradigmatique, habituellement virtuelle. D'une part, cette approche équivaut à introduire dans l'analyse une connaissance *a priori* de la langue étudiée, ce qui n'est pas compatible avec nos principes de travail ; d'autre part, elle rend les solutions proposées tributaires des structures linguistiques d'une langue donnée. C'est par exemple le cas de l'usage dans [RS97] de mots composés « redondants » (« *tuna fish* », « *oak tree* »), qui sont fréquents en anglais mais rares en français. De plus, nous ne disposons pas, dans le cadre de la sémantique différentielle, de principe interprétatif permettant de faire rigoureusement le lien entre ces motifs et les taxèmes de la théorie.

Pour toutes ces raisons, nous avons choisi d'adopter pour la construction de classes sémantiques l'approche illustrée par la première famille de travaux présentée dans cette section, faisant usage d'affinités du deuxième ordre. Pour

autant, divers impératifs nous éloignent des recherches existantes et imposent la mise au point de nouvelles méthodes adaptées aux spécificités de notre étude. Nous précisons ci-dessous ces points de divergence.

3.2.3. Positionnement relatif de nos recherches

Dans leur ensemble, les recherches que nous avons présentées dans la première partie de cette section font preuve d'une importante dépendance à l'analyse syntaxique — à l'exception de [HA94]et [MS96], qui ne réalisent cependant pas de classification sémantique à proprement parler, et de [SM99], qui ne considère qu'un sous-ensemble très codifié de la langue. Nous souhaitons comme nous l'avons déjà mentionné nous libérer de cette dépendance, ce qui est en théorie possible d'après la traduction en termes de sémantique interprétative du principe de similarité des contextes mentionné en introduction. Dans ce cadre théorique, en effet, deux contextes d'emploi de mots seront considérés comme ressemblants s'ils imposent sur leur « mot central » (*i.e.* le mot dont ils sont le contexte) des contraintes de sélection identiques ou similaires, impliquant que le mot central de l'un peut être employé dans l'autre, et réciproquement. Ces contraintes peuvent s'exprimer en termes d'obligation pour le mot de participer aux isotopies sémantiques présentes dans le contexte ; or, les isotopies sémantiques peuvent exister indépendamment de la syntaxe d'un énoncé, par simple rapport de proximité entre mots :

> « Une isotopie a une définition syntagmatique, mais non syntaxique ; elle n'est pas structurée, en d'autres termes, il s'agit d'un ensemble non ordonné. » ([Ras72], p. 82)

Travaillant ainsi à partir d'un concept linguistique potentiellement (bien que pas systématiquement) de « plus bas niveau » que la syntaxe, il est sans doute possible de caractériser le mode d'usage d'un mot dans un texte sans faire usage d'une analyse syntaxique de celui-ci — bien que cela soit sans doute au prix d'une perte d'information.

Un second point commun à la plupart des travaux présentés est la nécessité d'établir un compromis entre complexité de la langue étudiée et taille des données exploitées : quelle que soit la méthode employée, l'extraction non supervisée de classes sémantiques à partir de corpus ne relevant pas d'une langue de spécialité est limitée à des corpus de plusieurs dizaines de millions de mots au moins. Les résultats des recherches que nous avons évoquées dans cette section suggèrent que la constitution de classes sémantiques en employant des corpus généralistes de quelques centaines de milliers de mots, fussent-ils thématiquement homogènes, et sans connaissance sémantique *a priori*, est une tâche confinant à l'impossible.

Il semble en effet d'après les expériences que nous avons pu conduire dans ce sens que ce soit le cas. C'est pourquoi nous avons été amené à mettre au

point un système d'apprentissage en deux étapes. Dans un premier temps, nous calculons en employant le corpus intégral du *Monde diplomatique* une mesure évaluant la proximité sémantique entre mots ; cette mesure est naturellement assez fortement bruitée, notamment par les nombreux cas de polysémie présents dans ce corpus ni spécialisé, ni thématiquement cohérent. Elle constitue néanmoins un outil assez efficace pour enrichir, dans un second temps, l'analyse des contextes d'emploi des mots étudiés sur chacun des sous-corpus thématiques considérés. En effet, en autorisant la détection de similitudes sémantiques entre voisinages employant des mots distincts mais de sens proches, cette première évaluation de la proximité sémantique entre mots permet une généralisation des observations effectuées sur les données textuelles, et revient donc à « simuler » la disponibilité d'une quantité de données plus importantes. La section suivante commence par préciser les difficultés liées à la problématique abordée avant de décrire plus en détail le principe de travail que nous venons d'évoquer.

3.3. Approfondissement de la problématique et principe de la solution mise au point

Comme nous l'avons déjà précisé en introduction à ce chapitre, nous ne proposons pas ici de solution « parfaite » — si une telle chose peut exister — au problème de la génération automatique de classes sémantiques à partir de corpus. Notre travail de recherche nous a amené à explorer de nombreuses pistes, plus ou moins fructueuses et riches d'enseignements, dont nous ne décrivons en détail que la plus concluante. Il nous a néanmoins paru utile de faire état dans un premier temps (section 3.3.1) des réflexions d'ordre général sur la problématique traitée que ces travaux ont engendrées, afin de mettre en évidence les facteurs qui ont guidé nos choix. Nous présentons ensuite à la section 3.3.2 le principe général de la solution que nous proposons à ce problème, prenant en compte la spécificité de notre approche consistant à travailler sur des sous-corpus thématiquement cohérents extraits d'un corpus général de taille relativement importante.

3.3.1. Quelques réflexions sur les difficultés à surmonter

Nous introduisons ici plusieurs réflexions et précisions d'ordre général concernant la problématique traitée, puis procédons à quelques remarques plus « conjoncturelles » concernant les outils employés et conditions de réalisation de nos expériences.

3.3.1.1. Représentativité des données

La langue, et notamment telle qu'elle est employée dans notre corpus, permet d'envisager une quasi infinité (ou du moins, un nombre *a priori* indéfini) d'énoncés syntaxiquement valides et sémantiquement cohérents. Un corpus textuel de 11 millions de mots tel que celui que nous employons ne représente en conséquence qu'une part réellement minime des phénomènes langagiers imaginables — cela est d'autant plus vrai des sous-corpus tématiques que nous en avons extraits. Les travaux entrepris sur l'acquisition d'informations sémantiques à partir de corpus de langue générale (comme par exemple [LP01]) tendent à prouver qu'un corpus de 50 à 100 millions de mots commence à pouvoir être considéré comme statistiquement représentatif de la langue étudiée (cette taille pouvant diminuer jusqu'à 1 million, voir 100 000 mots dans le cas de langues de spécialité). Il résulte de cette simple constatation que le fait de n'avoir pas observé un énoncé dans nos données ne signifie en rien que cet énoncé soit impossible (ce qui est courant en analyse statistique de données) ni même qu'on puisse le considérer comme improbable (ce qui l'est moins). En d'autres termes, contrairement à Socrate dans la citation de Platon placée en exergue de ce chapitre, qui déduit une définition de « terrible » à partir d'exemples *et* contre-exemples d'usage, nous ne pouvons considérer nos données que comme sources d'informations *positives* concernant les potentialités d'emploi des mots. Il est nécessaire de prendre garde à cette propriété lors de la définition des mesures de similarité que nous employons : de nombreuses mesures statistiques évoluées tendent à prendre en compte l'absence d'une observation autant que sa présence. Bien qu'il soit impossible d'échapper totalement à ce biais, nous avons pu constater que les mesures statistiques les plus simples (voire rudimentaires), associées à des méthodes de normalisation spécifiquement mises au point pour s'adapter aux données étudiées, permettent souvent d'atteindre les résultats de classification les plus concluants.

3.3.1.2. Contexte et voisinage

Une conséquence de notre choix de ne pas réaliser d'analyse syntaxique du corpus d'étude est que nous nous trouvons contraint de simplifier la notion de « contexte d'apparition » d'un mot : ne pouvant le considérer comme un ensemble syntaxiquement structuré de mots entretenant, de par la formulation de l'énoncé étudié, une relation sémantique avec le mot-cible, nous devons l'approximer par la notion de *voisinage* d'une occurrence de mot. On désigne par là l'ensemble des mots apparaissant dans un rayon de n mots avant ou après une occurrence du mot-cible, sans franchir de ponctuation « forte »[2]. La question du mode de représentation adopté pour refléter l'information fournie par les voisinages est naturellement cruciale pour la recherche des similarités

[2]La définition d'une ponctuation forte peut être amenée à varier suivant les auteurs ; nous plaçons pour notre part dans cette catégorie tous les signes de ponctuation pouvant délimiter des phrases (« . », « ... », « ! », « ? »), ainsi que le point-virgule.

d'emplois entre mots, et nous aurons l'occasion d'y revenir au cours de la description du système proposé. Le choix d'une valeur de n optimale constitue un autre problème délicat, que nous approfondissons ci-dessous.

3.3.1.3. Taille de voisinage

La quantité de données devant être extraite du voisinage pour une occurrence donnée d'un mot de manière à en caractériser efficacement le sens constitue un problème difficile et toujours ouvert, et conditionne largement la qualité des résultats que l'on peut espérer obtenir. De nombreuses recherches ont abordé cette question, notamment dans le domaine de la désambiguïsation sémantique : D. Widdows [Wid03] développe un modèle mathématique du principe de l'association mot-sens guidée par le contexte, ce qui lui permet de définir la notion topologique de « groupe de contextes » comme l'ensemble des contextes tels qu'un mot donné exprime en leur sein le même sens, et de répondre de manière définitive à la question « Combien d'information contextuelle utiliser ? » par :

> « *Whatever is necessary.* »

Quoique d'un intérêt théorique indiscutable, ce travail reste néanmoins trop abstrait pour nous aider dans notre recherche. Plus concrètement, une étude systématique de la taille de voisinage la plus adaptée en désambiguïsation sémantique réalisée par L. Audibert [Aud03] suggère que le choix optimal pour cette tâche est en général de ne considérer qu'un ou deux mots pleins à droite et à gauche de la cible, cette valeur étant sujette à variations suivant notamment la catégorie du mot-cible.

La problématique est différente dans le cadre de la construction de classes sémantiques, notamment si la quantité de données disponible est relativement faible et implique de devoir en « tirer le maximum ». Un voisinage trop restreint ne permettra pas de rassembler suffisamment de mots indices pour bien caractériser la cible, notamment si le mot étudié apparaît peu souvent. Inversement, un voisinage trop grand risque d'amener l'analyse à prendre en compte de nombreux mots non pertinents — typiquement, ceci se traduit en pratique par l'obtention de classes sémantiques relevant plus de l'échelle du domaine (classes mésogénériques) que de celle du taxème (classes microgénériques).

Nous développons plus concrètement ce problème de la taille de voisinage lors de la description des expériences que nous avons menées, en montrant notamment la manière dont la prise en compte d'informations structurelles concernant la constitution des contextes permet de contourner en partie cette difficulté.

3.3.1.4. Pertinence du dénombrement des observations concomitantes

La stricte définition linguistique d'une classe sémantique suggère qu'il suffirait d'observer une seule fois deux mots employés dans un même contexte pour pouvoir les regrouper au sein d'une même classe. Néanmoins, tous les contextes « ne se valent pas », et certains exercent une contrainte beaucoup plus forte que d'autres concernant le sens du mot tenant la place du mot cible : ainsi dans la phrase « Lauteur vient de sortir chez Léditeur un ouvrage remarquable, X, qu'il dédicacera le 11 mars à la librairie Lelibraire », 19 mots de contexte réussissent l'exploit consistant à ne donner aucune information de quelque sorte que ce soit sur X. Il n'est malheureusement pas envisageable sans connaissance sémantique *a priori* de proposer un ou plusieurs critères pour évaluer de manière automatique cette « pertinence de voisinage ». Nous ne pouvons donc tirer de confiance dans la similarité sémantique de deux mots que par le nombre de voisinages qui les rapprochent.

Est-il en revanche pertinent de prendre en compte le nombre de fois où un même voisinage rapproche deux mots ? En théorie, une observation suffit pour affirmer que deux mots sont interchangeables dans un contexte donné, d'autres observations ne faisant que confirmer cette possibilité. Il peut pourtant être important de savoir si le contexte commun observé correspond ou non à un usage important (en volume) des deux mots comparés. Les expériences que nous avons menées suggèrent d'adopter à cet égard une position différente suivant la quantité de données disponible. Si celle-ci est importante, il semble plus pertinent de dénombrer les éléments de contexte distincts communs aux deux mots, afin d'éviter notamment les biais pouvant être introduits par les termes complexes, où les mots peuvent être employés dans un sens « déviant ». Si elle est faible, il est en revanche important de prendre en compte toute l'information disponible, fût-elle biaisée, et donc de considérer pour la comparaison des usages de mots les nombres d'occurrences de mots-indices dans leurs voisinages.

3.3.1.5. Variation du volume de données

Une difficulté importante rencontrée au cours de notre recherche est la variation du volume de données disponible pour construire la représentation de chaque mot étudié. Il est fréquent qu'au cours de la classification d'une sélection de mots, le rapport entre les nombres d'occurrences des mots le moins fréquent et le plus fréquent soit du simple au centuple.

Pour la plupart des mesures de similarité « évoluées » que nous avons développées, il n'est pas possible de trouver un facteur de normalisation simple dépendant des seuls nombres d'occurrences des mots comparés. Nous avons donc mis au point deux techniques originales permettant de calculer malgré tout des similarités entre mots relativement indépendantes de leurs nombres d'occurrences respectifs : la première est une méthode mathématique de nor-

malisation *a posteriori* des valeurs de similarité rassemblées dans une matrice, tandis que la seconde évalue la similarité entre deux mots à partir des similarités calculées entre deux populations de mots « factices » présentant tous le même nombre d'occurrences, générés par échantillonnage aléatoire des voisinages caractérisant les mots comparés. Ces deux techniques numériques sont décrites dans ce chapitre à mesure que leur usage est requis par les méthodes de construction de classes sémantiques développées.

3.3.1.6. Conditions expérimentales

Sauf indication contraire, les expériences présentées dans ce chapitre ont été réalisées en employant les données textuelles du sous-corpus construit au chapitre précédent consacré aux « nouvelles technologies » (correspondant au domaine désigné par le triplet de mots < informatique / réseau / communication >). Ce sous-corpus a été sélectionné pour le volume de données relativement important qu'il rassemble (environ 400 000 mots), pour la richesse des sens spécifiques que les mots sont susceptibles d'y développer, ainsi qu'à cause de la confiance que nous avions de pouvoir évaluer sans erreur la pertinence des résultats pour ce thème. Les mots étudiés pour la classification sont sélectionnés selon un critère de fréquence parmi ceux du domaine correspondant.

Nos recherches ont principalement porté sur la définition d'une mesure de similarité entre mots qui reflète aussi fidèlement que possible la proximité sémantique existant entre eux. Les valeurs de similarité ainsi générées sont exploitées par un algorithme de classification ascendante hiérarchique simple faisant usage du lien moyen pour évaluer la similarité entre classes (*i.e.* réalisant pour cela la moyenne des similarités entre objets des deux classes). Le choix de ne pas employer une fois encore la méthode CHAVL utilisée au chapitre précédent résulte principalement de notre volonté de pouvoir évaluer sans biais la qualité des mesures de similarité développées : si l'arbre de classification propose des regroupements pertinents, il est dans ces conditions indiscutable que c'est le fait d'une « bonne » mesure de similarité, et non pas de la sophistication de l'algorithme de classification employé.

Si nous avons réussi à faire apparaître des classes homogènes dans les arbres de classification construits, nous ne sommes en revanche pas parvenu à mettre au point pour la problématique étudiée dans ce chapitre une technique permettant de les en extraire automatiquement, à l'instar de la méthodologie proposée dans ce sens pour la constitution de classes de mots-clés thématiques au chapitre 2. L'évaluation des résultats obtenus est donc effectuée en extrayant à la main d'un arbre de classification proposé les classes jugées comme sémantiquement pertinentes, et à calculer quelle proportion de la totalité des mots classés elles rassemblent.

Ces précisions d'ordre divers ayant été apportées, nous pouvons désormais procéder à la présentation des principes généraux ayant guidé notre recherche.

3.3.2. Structure des travaux menés

Il est apparu rapidement au cours de nos expériences que le principal obstacle à une comparaison efficace des voisinages d'occurrences des mots étudiés est le phénomène de variabilité lexicale des énoncés : deux contextes peuvent ainsi être sémantiquement très proches et exercer des contraintes similaires sur les mots dont ils constituent l'entour linguistique sans pour autant avoir aucun mot en commun. Une approche directe de la comparaison des modes d'usages de deux mots, étudiant les voisins apparaissant dans leurs contextes, est ainsi condamnée à souffrir d'une perte d'information considérable, et à ne pouvoir exploiter pour rapprocher les mots considérés qu'une faible partie des indices effectivement présents dans le texte. Cette perte est compensée dans le cas de corpus de grande taille par la quantité de données disponibles, mais s'avère réellement problématique pour l'exploitation de corpus de petite taille. Une solution à cette difficulté consiste à faire usage de premières classes sémantiques afin de réduire l'impact de cette variabilité en généralisant les observations effectives — c'est ce que fait Z. Harris lorsque, dans l'exemple donné en introduction à ce chapitre, il réinvestit dans l'analyse des énoncés étudiés la connaissance de l'équivalence « lune-lac » afin de pouvoir en déduire l'équivalence « sale-rouge ». Puisque nous refusons par principe tout apport de connaissances auxiliaires, il ne nous est pas possible de faire appel à une base de données externe (par exemple un thesaurus prédéfini) susceptible de nous fournir cette première organisation en classes du vocabulaire observé. C'est pourquoi nous réalisons la construction des classes sémantiques selon les deux étapes déjà mentionnées : une mesure approximative de similarité sémantique entre mots est calculée en employant l'intégralité des données du corpus du *Monde diplomatique*, et réinvestie dans l'étude des voisinages de mots pour la construction de classes sémantiques sur les sous-corpus thématiques. Ces deux tâches soulèvent des problématiques assez différentes, que nous précisons maintenant.

Dans un premier temps, notre objectif est de réaliser une classification sémantique des mots sur l'ensemble du corpus, sans aucune préconnaissance sémantique, et en ne sélectionnant les mots que selon un critère de fréquence (le découpage en domaines est ignoré). L'impératif est ici de pouvoir traiter de manière raisonnablement efficace une grande quantité de données ; de plus, la représentation de ces données se doit d'être assez tolérante afin de faire apparaître un maximum d'indices positifs justifiant le rapprochement de mots. La solution proposée, présentée à la section 3.4, consiste à utiliser une caractérisation ensembliste regroupant de manière non structurée pour chaque mot à classer tous les mots susceptibles d'entretenir dans le texte un lien syntagmatique avec lui. La mesure de similarité employée est directement adaptée du

classique indice de Jaccard*. Nous aboutissons de cette manière à une classification des mots suffisamment fiable pour être exploitée par la suite.

Dans un second temps, nous construisons en employant les données textuelles rassemblées dans les sous-corpus thématiques, des classes sémantiques à l'intérieur des domaines correspondants, en tirant profit des connaissances acquises lors du traitement précédent. La problématique est alors de tirer le maximum d'information des données disponibles, et d'éviter les faux indices positifs, qui ne sont plus dans cette nouvelle configuration « lissés » par le volume de données. Nous proposons à la section 3.5 pour satisfaire cette contrainte une méthodologie originale consistant à définir non pas une représentation synthétique de l'ensemble des voisinages d'un mot accompagnée d'une mesure de similarité *ad hoc*, mais une mesure de similarité entre voisinages individuels fondée sur la mesure de proximité entre mots calculée à l'étape précédente, accompagnée d'un mécanisme permettant le calcul d'une similarité entre mots à partir des similarités ainsi mesurées entre leurs voisinages.

3.4. Classification sémantique des mots sur l'ensemble du corpus

Pour rester cohérent avec nos partis pris théoriques, nous ne pouvons rigoureusement considérer les techniques présentées dans ce chapitre comme répondant à la question de construction de classes sémantiques : elles exploitent en effet sans sélection la totalité des données textuelles rassemblées dans notre corpus d'archives du *Monde diplomatique*, et ignorent le découpage du lexique en domaines réalisé à l'étape précédente. Pour autant, les méthodes employées se fondent sur les mêmes principes d'évaluation de la similarité d'usage des mots, et les résultats obtenus esquissent bien des regroupements qualifiables, à ce *distinguo* théorique près, de classes sémantiques.

Afin de favoriser l'exploitation ultérieure de la mesure de proximité sémantique entre mots calculée dans cette section, nous souhaitons que celle-ci puisse s'interpréter comme une distance, vérifiant donc les propriétés de symétrie et d'inégalité triangulaire. Plutôt que d'imposer ces contraintes à la mesure employée pour comparer les représentations de mots elle-même, le système mis en œuvre utilise cette première mesure de similarité « brute » afin de construire un arbre de classification, sur lequel nous définissons une métrique simple. C'est cette métrique qui est ensuite employée comme préconnaissance lors de la construction de classes sémantiques sur des sous-corpus thématiques.

Nous présentons dans un premier temps la représentation et la méthode de sélection des indices contextuels mises au point pour associer à chaque mot à classer une caractérisation de l'ensemble des voisinages dans lesquels il est employé (section 3.4.1), puis la mesure de similarité définie entre ces caractérisations (section 3.4.2) ainsi que la méthode statistique élaborée afin de norma-

liser celle-ci pour optimiser son pouvoir discriminant (section 3.4.3). La section 3.4.4 est ensuite consacrée à l'étude des classifications de mots que cette mesure nous permet d'engendrer, avant que nous ne revenions sur un problème soulevé par la technique décrite : l'exploitation des voisinages d'apparition des mots classés afin d'en extraire des indices contextuels caractérisant leurs usages est guidée par quelques règles qui, quoique très simples, s'apparentent à une ébauche d'analyse syntaxique de surface et remettent en cause l'indépendance de nos travaux vis-à-vis de la langue étudiée. C'est pourquoi nous présentons en section 3.4.5 une variante du mode de représentation des voisinages proposé dans un premier temps, qui permet d'éviter cet écueil, au prix toutefois d'une légère baisse de qualité des résultats.

3.4.1. Principe général

La méthode introduite ici est très directement inspirée des divers travaux présentés en section 3.2.1, mais ne fait pas explicitement usage d'une analyse syntaxique : chaque mot étudié est caractérisé par un ensemble regroupant les mots apparaissant « à une bonne position » (précisée ci-dessous) dans au moins n voisinages de ses occurrences. Le paramètre n permet de se limiter aux mots dont il est relativement sûr que leur apparition dans le voisinage du mot-cible n'est pas absolument fortuite ; nous avons empiriquement établi sa valeur à $n = 2$, ce qui permet un « filtrage » élémentaire au prix d'une perte d'information minimale.

Les deux principaux facteurs susceptibles de faire varier les résultats obtenus sont d'une part la taille des voisinages considérés, et d'autre part le mode de sélection des mots pris en compte pour la classification en fonction de leur catégorie grammaticale et de leur position par rapport au mot-cible. Nous listons ci-dessous les paramètres de sélection optimaux obtenus à la suite de nombreuses expériences, en fonction de la catégorie morphosyntaxique des mots à classer[3].

Adjectifs

- taille de voisinage : 1 mot à droite et à gauche,
- types retenus : noms et adverbes.

Adverbes

- taille de voisinage : 1 mot à droite et à gauche,
- types retenus : verbes et adjectifs.

[3] Étant donné l'usage devant être fait des résultats de cette section, nous nous intéressons à la classification des principales catégories de mots pouvant servir d'indices lors de la comparaison des contextes d'usages de deux mots : noms et noms propres, adjectifs, adverbes, verbes et nombres — ces derniers permettant d'identifier et caractériser les noms d'unités ou, de manière plus intéressante, les noms « désignés en unités » : « n tonnes de { blé, riz, maïs }… ».

Nombres

- taille de voisinage : 1 mot à gauche, 2 à droite,
- types retenus : noms immédiatement contigus à gauche, et immédiatement contigus ou séparés par la préposition *de* à droite (afin de prendre en compte les « millions de », *etc.*).

Noms (et noms propres)

- taille de voisinage : 3 mots à droite et à gauche,
- types retenus : verbes, noms propres (dont les acronymes), adjectifs immédiatement contigus à gauche ou à droite, prépositions apparaissant à gauche si aucun autre nom n'est intercalé, noms apparaissant à gauche ou à droite si une préposition est intercalée, nombres apparaissant à gauche.

Verbes

- taille de voisinage : 3 mots à droite et à gauche,
- types retenus : verbes et adverbes immédiatement contigus, noms et noms propres (dont les acronymes).

Chacun des mots à classer est ainsi caractérisé par un ensemble de mots sélectionnés de manière à maximiser, autant qu'il est possible de le faire sans réaliser d'analyse syntaxique du texte, la probabilité d'existence d'un lien sémantique entre eux et le mot cible. Nous nous penchons maintenant sur la mesure de similarité ensembliste employée afin de comparer ces représentations.

3.4.2. Mesure de similarité

La représentation ensembliste des mots que nous venons de présenter a donné lieu à de nombreuses expériences de classification employant des mesures de comparaison entre ensembles d'attributs binaires plus ou moins classiques et évoluées (en particulier les mesures incorporées au programme `chavleps` que nous employons au chapitre 2). La plupart de celles-ci ne permettent d'obtenir que des résultats de qualité moyenne, notamment du fait du biais mentionné à la section 3.3.1.4 consistant à considérer l'absence partagée d'un attribut comme un indicateur de similarité aussi probant que sa présence. Les résultats les plus satisfaisants sont en définitive obtenus grâce à un simple indice de Jaccard normalisé par une méthode que nous avons mise au point dans le but de limiter les effets de « masse de données » mentionnés en section 3.3.1.5. On constate en effet en réalisant une classification ascendante hiérarchique employant un indice de Jaccard brut que certains mots, notamment parmi les plus fréquents, tendent à présenter des valeurs de similarité relativement fortes avec presque tous les autres ; cela induit sur la classification ce que l'on pourrait nommer un « effet trou noir », où tous les objets à classer s'agglutinent autour d'un même noyau de classe très dense au lieu de former des classes clairement distinctes dans l'arbre. Le principe de la méthode mise

au point pour éviter cet effet est de normaliser la matrice de similarité afin de ramener dans une même échelle de valeurs toutes les lignes et colonnes de celle-ci. Nous détaillons dans ce qui suit les calculs mis en œuvre afin de procéder à cette normalisation.

3.4.3. Normalisation *a posteriori* d'une matrice de similarité

L'objet des traitements présentés ici est une matrice $\mathcal{M}$ rassemblant les similarités entre paires d'objets à classer ; cette matrice est symétrique, et volontiers représentée dans la littérature consacrée à l'analyse statistique de données sous la forme d'une demi-matrice. Nous choisissons néanmoins ici par commodité de considérer la matrice symétrique dans son ensemble ; les modifications que nous nous proposons de lui apporter devront naturellement maintenir cette propriété de symétrie.

Il est courant, lors de l'analyse de données statistiques, de transformer une matrice de similarité afin de la rendre « centrée et réduite ». Cette opération consiste en une translation des valeurs rassemblées dans la matrice ramenant leur moyenne générale à 0 (centrage), suivie d'une mise à l'échelle amenant leur écart type à 1 (réduction). Nous présentons ici une transformation inspirée de cette méthode classique, mais visant à centrer et réduire *chaque ligne* (et chaque colonne) de la matrice de similarité, tout en maintenant la symétrie de celle-ci. Cette opération, qui constitue à notre connaissance une contribution originale dans ce domaine, nous garantit que les similarités entre chaque objet et l'ensemble des autres soient du même ordre de grandeur et évoluent dans la même « plage de valeur ». Le bénéfice attendu en est très important lorsqu'au cours d'une classification certains objets ont tendance à se comporter comme « attracteurs » parce qu'ils possèdent de nombreuses caractéristiques communes à beaucoup d'objets.

Centrage La première opération réalisée consiste à centrer les lignes et colonnes de la matrice en maintenant sa propriété de symétrie. Si l'on note c_i le décalage appliqué à tous les éléments de la i-ème ligne de la matrice afin de ramener leur moyenne à 0, celui-ci devra, afin de maintenir cette symétrie, être également appliqué à sa i-ème colonne. La matrice $\mathcal{M}' = \left(m'_{ij}\right)$ résultant de l'opération de centrage ligne par ligne de $\mathcal{M} = (m_{ij})$ est donc définie par :

$$\forall (i,j) \in \{1..n\}^2 \;\; m'_{ij} = m_{ij} + c_i + c_j \tag{3.1}$$

où n est le nombre de lignes (et colonnes) de $\mathcal{M}$.

Les m'_{ij} vérifient, par définition :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m'_{ij} = 0 \quad \text{(centrage global)} \tag{3.2}$$

$$\forall i \in \{1..n\} \sum_{j=1}^{n} m'_{ij} = 0 \quad \text{(centrage des lignes)} \tag{3.3}$$

L'équation 3.2 traduit la propriété de centrage global de la matrice, qui est une conséquence directe du centrage de chaque ligne ; exprimée en termes de m_{ij} et c_i, elle devient :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} + c_i + c_j = 0 \tag{3.4}$$

On en déduit :

$$n^2\bar{m} + n^2\bar{c} + n^2\bar{c} = 0, \text{ soit } : \bar{c} = -\frac{\bar{m}}{2} \tag{3.5}$$

où $\bar{m}$ désigne la moyenne de tous les m_{ij} et $\bar{c}$ celle des c_i.

L'équation 3.3 traduit la propriété de centrage de chaque ligne de la matrice prise individuellement (et donc de chaque colonne, la matrice étant symétrique). Elle se développe en :

$$\forall i \in \{1..n\} \; nc_i + n\bar{m}_i + n\bar{c} = 0 \tag{3.6}$$

où $\bar{m}_i$ désigne la moyenne des valeurs de la i-ème ligne de la matrice $\mathcal{M}$. De 3.5 et 3.6, on déduit :

$$\forall i \in \{1..n\} \; c_i = \frac{\bar{m}}{2} - \bar{m}_i \tag{3.7}$$

Il est ainsi possible d'obtenir simplement la matrice $\mathcal{M}'$ correspondant à $\mathcal{M}$ mais dont toutes les lignes (et colonnes) sont centrées, c'est-à-dire de moyenne nulle.

Réduction De même que nous avons pour la tâche de centrage des lignes de la matrice calculé pour chaque ligne une translation c_i, nous calculons maintenant les valeurs des k_i, coefficients multiplicateurs à appliquer aux lignes et colonnes de $\mathcal{M}'$ pour obtenir $\mathcal{M}''$, toujours symétrique mais dans laquelle l'écart type (et donc la variance) de chaque ligne et colonne est égal à l'unité. Comme lors du calcul précédent, si le coefficient k_i est appliqué à toutes les valeurs de la i-ème ligne de $\mathcal{M}'$, alors il doit également être appliqué à sa i-ème colonne, afin de maintenir la symétrie de la matrice. On a donc :

$$\forall (i,j) \in \{1..n\}^2 \; m''_{ij} = k_i k_j m'_{ij} \tag{3.8}$$

Grâce à la propriété de centrage de la matrice $\mathcal{M}'$, l'équation exprimant la réduction de l'écart type de chaque ligne (et colonne) de $\mathcal{M}''$ se simplifie en :

$$\forall i \in \{1..n\} \qquad \sum_{j=1}^{n} \left(m''_{ij}\right)^2 = n \tag{3.9}$$

$$\sum_{j=1}^{n} \left(k_i k_j m'_{ij}\right)^2 = n \tag{3.10}$$

La forme de ce système d'équations ne se prête malheureusement pas à une résolution directe ; néanmoins, il découle de 3.10 que :

$$k_i = \sqrt{\frac{n}{\sum_{j=1}^{n} \left(k_j m'_{ij}\right)^2}} \tag{3.11}$$

L'équation 3.11 fournit une base raisonnablement efficace pour un calcul itératif des k_i : nous proposons de définir pour tout i la suite (k_i) par :

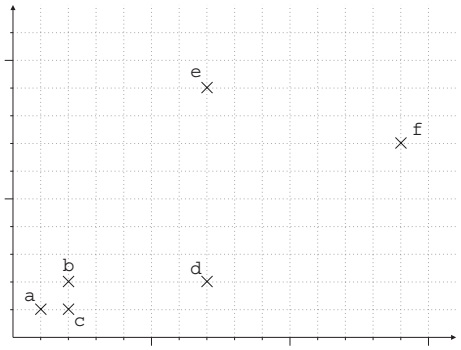
$$\begin{cases} k_{i\,0} & = & 1 \\ k_{i\,(n+1)} & = & \frac{1}{2}\left(k_{i\,n} + \sqrt{\frac{n}{\sum_{j=1}^{n}\left(k_{j\,n}.m'_{ij}\right)^2}}\right) \end{cases} \tag{3.12}$$

Le choix de définir $k_{i\,(n+1)}$ comme la moyenne de $k_{i\,n}$ avec la nouvelle valeur calculée par la formule de l'équation (3.11) nous permet d'éviter les « oscillations » de la suite et d'assurer (du moins empiriquement) sa convergence. Pour les tailles de données usuelles (de l'ordre du millier d'objets classés), la formule récursive donnée par l'équation (3.12) permet le calcul de valeurs approchées à 10^{-6} près des k_i en une vingtaine d'itérations.

L'application à la matrice $\mathcal{M}'$ des coefficients de mise à l'échelle k_i présente bien sûr l'inconvénient de lui faire perdre sa propriété de centrage. C'est pourquoi nous procédons en pratique à la répétition des deux procédés de centrage et réduction ligne par ligne jusqu'à stabilisation à ε près des valeurs de la matrice, ce qui se produit en général, suivant la rigueur de l'estimation souhaitée, après quelques dizaines d'itérations.

La figure 3.1 fournit un exemple d'application de la technique que nous proposons à un cas simple de classification de six objets définis par leurs coordonnées dans un espace euclidien à deux dimensions. On constate qu'en amenant les mesures de similarité de tous les objets dans la même plage de valeurs, elle réalise une sorte de « distorsion » de l'espace selon laquelle la similarité entre deux objets ne dépend plus directement de la distance qui les sépare, mais de l'importance de celle-ci comparativement aux distances qu'ils entretiennent avec les autres objets.

Soit six objets définis par leurs coordonnées dans l'espace euclidien ordinaire :



$a\binom{1}{1}$ $b\binom{2}{2}$ $c\binom{2}{1}$

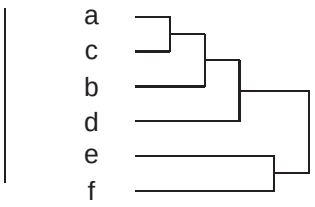
$d\binom{7}{2}$ $e\binom{7}{9}$ $f\binom{14}{7}$

Soit une mesure de similarité entre ces objets, fonction de la distance qui les sépare :

$$s(o_1, o_2) = \frac{1}{1 + \|o_1, o_2\|}$$

Ses valeurs approximatives sont données par la demi-matrice $\mathcal{M}$ suivante :

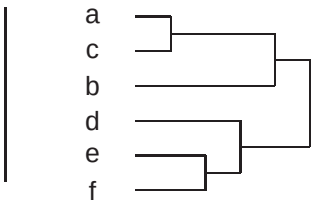
	a	b	c	d	e	f
a	1					
b	0.414	1				
c	0.500	0.500	1			
d	0.141	0.167	0.164	1		
e	0.091	0.104	0.096	0.125	1	
f	0.065	0.071	0.069	0.104	0.121	1



Nous avons fait figurer en regard de la matrice de similarité l'arbre de classification résultant de son exploitation par un algorithme de classification ascendante hiérarchique simple par lien moyen. On constate que celle-ci réalise l'agrégation de d avec la classe a, b, c, ce qui se justifie d'un point de vue purement numérique (d est plus proche de b que de e) mais est peu satisfaisant si l'on veut faire apparaître une organisation « logique » globale des objets.

Nous présentons maintenant l'état de la matrice après 20 itérations de l'algorithme de centrage et réduction ligne par ligne présenté à la section 3.4.3 (à ce stade, toutes les lignes sont réduites et centrées à 10^{-8} près), en faisant similairement figurer la classification induite :

	a	b	c	d	e	f
a	2.153					
b	−0.230	2.160				
c	0.060	0.027	2.098			
d	−0.660	−0.616	−0.688	2.166		
e	−0.664	−0.662	−0.750	−0.106	2.132	
f	−0.658	−0.679	−0.746	−0.096	0.051	2.130



On observe cette fois la création de deux classes a, b, c et d, e, f, toutes deux homogènes du point de vue des similarités entre leurs composants.

FIG. 3.1 – *Illustration de l'intérêt du procédé de centrage et réduction ligne par ligne et colonne par colonne d'une matrice de similarité*

Il est à noter que la méthode de l'AVL d'I. C. Lerman, que nous employons au chapitre 2 (section 2.3.3), est également à même de traiter « intelligemment » le type de situation présenté à la figure 3.1, pour un coût calculatoire moindre et de manière sans doute mieux fondée statistiquement, mais nous avons été contraint par souci de clarté de donner un exemple relativement simple qui ne reflète pas totalement les difficultés qui nous ont amené à développer la technique proposée.

3.4.4. Résultats

La méthode relativement simple d'évaluation des proximités d'usage entre mots présentée dans cette section nous permet de réaliser par analyse des 11 millions de mots du corpus complet des regroupements de mots en arbres de classification sémantique de qualité assez satisfaisante. Comme précisé à la section 3.3.1.6, la méthode employée pour réaliser la classification est volontairement simple ; en particulier, elle ne permet pas le multiclassage, et nous n'avons pas prêté d'attention particulière au cas des mots polysémiques, qui sont dans le meilleur des cas rattachés à une classe correspondant à un de leurs sens. Les figures 3.2 et 3.3 présentent des extraits des arbres de classification obtenus pour tous les noms et adjectifs, respectivement, apparaissant plus de 100 fois dans le corpus — il semble expérimentalement que cette valeur représente le nombre minimal d'occurrences en-deçà duquel la caractérisation des mots par l'ensemble de leur voisins dans le texte n'est plus suffisamment discriminante pour permettre une classification de qualité acceptable ; nous discutons ci-après plus en détail de l'impact du choix de ce nombre d'occurrences minimal sur la qualité des résultats.

Les arbres produits permettent en moyenne le regroupement en classes pertinentes (par sélection manuelle de celles-ci dans l'arbre) des deux tiers des mots classés. Nous donnons ici une description rapide des résultats obtenus qui devrait, avec les deux figures déjà citées, permettre au lecteur de s'en faire une idée assez précise. L'annexe B, page 185 présente une sélection plus complète de classes extraites manuellement des arbres de classification construits.

Noms

C'est pour cette catégorie que l'on observe les rapprochements les plus systématiquement pertinents. Sont par exemple regroupés des indicateurs de date et durée, des institutions, des acteurs / professions, ou des tendances (*augmentation*, *accroissement*, *etc*).

Adjectifs

Des classes d'ordinaux, de nationalités, d'indicateurs de taille ou d'importance, de couleurs, ou de tendances politiques sont par exemple clairement identifiées. La classification est dans son ensemble d'une qualité inférieure aux autres en raison d'erreurs lors de l'étiquetage du corpus : de nombreux participes de verbes ont ainsi été désignés comme adjectifs, alors que leur contexte ne « fait pas sens » pour caractériser un adjectif.

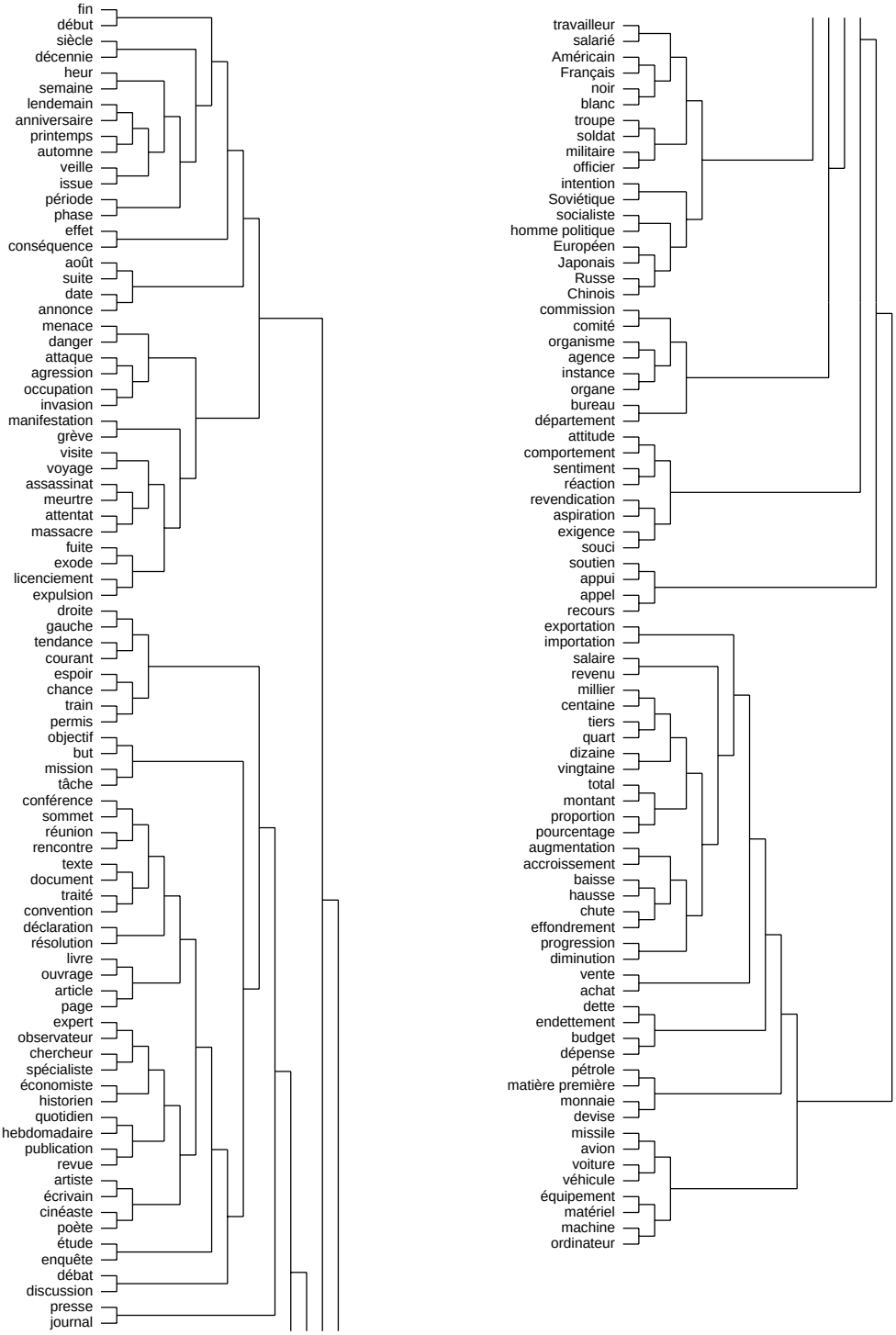


FIG. 3.2 – *Sous-arbre extrait de l'arbre de classification sémantique des noms apparaissant plus de 100 fois dans le corpus, obtenu par la méthode « globale » prenant en compte l'intégralité des données du corpus*

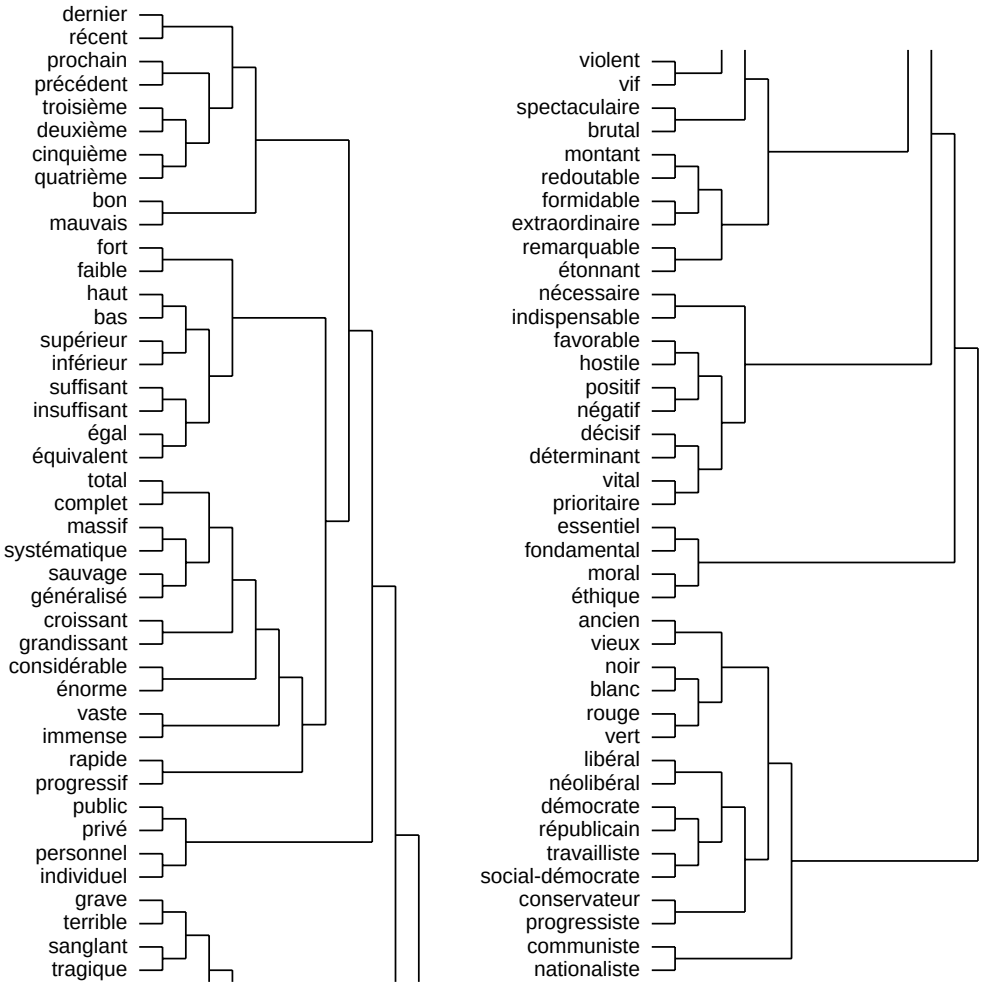


FIG. 3.3 – *Sous-arbre extrait de l'arbre de classification sémantique des adjectifs apparaissant plus de 100 fois dans le corpus, obtenu par la méthode « globale » prenant en compte l'intégralité des données du corpus*

Adverbes

Moins nombreux que les précédents à seuil de nombres d'occurrences égal, ils forment des classes moins nettement définissables ; on remarque néanmoins les adverbes d'intensification (*fort*, *particulièrement*, *profondément*, *etc.*), d'approximation ou de temps (*tôt*, *tard*, *etc.*), ainsi que de nombreuses paires pertinentes mais isolées ({ *publiquement*, *ouvertement* }, { *en effet*, *pourtant* }).

Verbes

Les indicateurs de modalité (*devoir*, *falloir*, *pouvoir*, *etc.*) se rapprochent des verbes « de conséquence » — *provoquer*, *entraîner*, *etc.* —, l'ensemble se détachant nettement du reste. De manière peu surprenante pour un

corpus journalistique, une volumineuse classe est consacrée à l'expression (*affirmer*, *déclarer*, *admettre*, *souligner*, *manifester*, *témoigner*, *etc.*) ; on peut remarquer le rapprochement de paires de mots de sens réciproques (*expliquer* / *comprendre*), ce qui s'explique notamment par le fait que la caractérisation mise au point mélange dans le même ensemble les mots du voisinage précédent et suivant le mot cible.

Nombres

Les nombres (beaucoup moins nombreux que les précédents à seuil de nombres d'occurrences égal) se répartissent pour leur part en dates (typiquement, $19xx$), « petits nombres » (écrits en toutes lettres), nombres « imposants » (*million*, *milliard*) destinés à la désignation de sommes d'argents, nombres décimaux (pourcentages), et nombres « ordinaires », principalement employés dans le corpus pour la désignation des pages.

La qualité globale des arbres de classification produits présente une dépendance assez forte au nombre d'occurrences minimal guidant le choix des mots classés, d'une part parce qu'en abaissant celui-ci on augmente le nombre d'objets étudiés, ce qui rend naturellement leur organisation « intelligente » plus délicate, et d'autre part parce que la quantité de données disponibles pour caractériser chaque mot dépend également de cette valeur. Expérimentalement, nous avons pu constater que les arbres de classification obtenus sur les mots apparaissant plus de 500 fois dans le corpus sont de bonne, voire très bonne qualité. Pour un seuil de 200 occurrences au minimum, ils restent relativement bons, la principale perte de qualité étant que, bien que la plupart des rapprochements entre mots effectués soient pertinents, tous les rapprochements pertinents ne sont pas effectués : certains groupes de mots se trouvent ainsi scindés en plusieurs sous-groupes séparés dans l'arbre. En abaissant le seuil à 100 occurrences, on constate que l'arbre obtenu à 200 occurrences reste presque inchangé : les nouveaux mots introduits dans la classification forment de très petites classes sémantiquement justifiées en elles-mêmes mais qui s'agglutinent de manière désordonnée « par-dessus » l'arbre de classification des mots plus fréquents, au lieu de s'intégrer à celui-ci. Quoique l'arbre obtenu dans ces dernières conditions soit moins « intéressant », notre objectif est de définir une première mesure de proximité entre mots qui couvre une part aussi grande que possible du vocabulaire du corpus ; c'est pourquoi nous retenons pour la suite ce seuil de 100 occurrences comme valeur de travail.

Remarquons enfin que le succès de la méthode proposée est tributaire de son application au corpus complet du *Monde diplomatique* : appliquée à titre d'essai à une sélection de 300 noms du corpus « nouvelles technologies », elle permet de générer un arbre dans lequel moins d'une vingtaine de mots sont classés correctement, et les autres ne présentent même pas d'organisation globale reconnaissable. S'il est compréhensible que la qualité de la classification générée baisse de manière importante avec la diminution de la quantité de données exploitables, la chute observée semble plus importante que ce seul

phénomène ne peut en rendre compte. Une seconde explication possible est la plus grande diversité de données textuelles rassemblée par le corpus complet comparativement aux sous-corpus thématiques. Dans le corpus intégral, en effet, chaque mot est susceptible d'être employé dans des contextes très variés, en particulier suivant le domaine où il est observé ; cette variété présente le double avantage d'augmenter la probabilité de pouvoir observer des points communs avec d'autres mots et d'accentuer le « relief » de la représentation construite pour le mot étudié, et donc son pouvoir discriminant. Dans le cas des sous-corpus thématiques, l'usage des mots est plus susceptible d'être « normalisé » et restreint, limitant les « aspérités » nourrissant le pouvoir discriminant de sa caractérisation. Il se confirme donc qu'il est nécessaire de mettre au point des techniques particulières, analysant plus finement les données contextuelles, pour la construction de classes sémantiques à partir des sous-corpus thématiques.

Afin de permettre l'exploitation des résultats obtenus, qui constituent une première structuration généraliste relativement pertinente de l'espace sémantique, nous définissons à partir des arbres de classification ainsi produits une distance reflétant leur structure : la distance entre deux mots a et b est définie comme le logarithme à base 2 du cardinal de la plus petite classe de l'arbre qui les rassemble (en d'autres termes, le nombre de feuilles du plus petit sous-arbre de l'arbre de classification dont a et b sont des feuilles). Il s'agit, comme toutes les distances définies sur de tels arbres, d'une ultramétrique*, c'est-à-dire d'une métrique d vérifiant une forme forte de l'inégalité triangulaire :

$$\forall(a,b,c) \quad d(a,b) \leq \max\left(d(a,c), d(b,c)\right)$$

ce qui ne réduit en rien sa qualité de distance. La formule choisie utilisant le $\log_2$ est équivalente à la hauteur dans l'arbre à laquelle se trouverait le nœud effectuant la fusion permettant de rassembler les deux mots comparés si l'arbre était binaire et équilibré — il s'agit ainsi d'une manière simple de refléter de manière numérique l'information structurelle représentée par l'arbre. Les valeurs ainsi obtenues sont normalisées à 1 afin d'être exploitables indifféremment comme distances ou similarités (par différence à 1).

La méthode que nous venons de décrire serait tout à fait satisfaisante pour notre besoin immédiat si la procédure permettant la sélection des mots permettant de caractériser l'usage des mots classés ne pouvait sembler obéir à des critères arbitraires et spécifiques aux constructions, sinon d'une langue unique, du moins d'une famille de langues. Bien que des impératifs concernant la qualité des résultats nous imposent finalement d'employer dans la suite la technique déjà décrite, nous introduisons à la section suivante, « entre parenthèses », une solution alternative pouvant permettre si nécessaire d'éviter ce biais.

3.4.5. Alternative : représentation « ensembliste typée » des voisinages

Cette variante de la technique précédente a été développée afin d'explorer les possibilités d'extension multilingue du système : le schéma de sélection des mots indices décrit à la section 3.4.1 est en effet guidé par des considérations d'ordre syntaxique, assez spécifiques à la langue étudiée. Nous proposons donc une méthode permettant d'obtenir des résultats comparables à la précédente, quoique légèrement moins précis, sans requérir de tels présupposés sur les constructions de la langue dans laquelle le texte étudié est rédigé.

3.4.5.1. Principe

De même que pour le mode de représentation décrit précédemment, nous associons à chaque mot étudié l'ensemble des mots apparaissant au voisinage d'une de ses occurrences, mais sans faire cette fois usage de règles permettant d'effectuer une sélection aussi pertinente que possible : les deux seules critères sont cette fois les catégories syntaxiques des mots susceptibles de jouer le rôle d'indices, limitées aux noms, verbes, adjectifs, adverbes, nombres et noms propres, et la taille de voisinage considéré, fixée à 4 mots avant ou après le mot cible pour nos expériences (comme nous le voyons plus loin, cette valeur n'a ici que peu d'importance).

En revanche, chacun des mots indices ainsi extrait des voisinages du mot à caractériser reste associé dans la représentation construite à la séquence des catégories grammaticales des mots qui le séparent du mot cible (en ignorant les « déterminants » : articles, adjectifs démonstratifs, *etc.*). Pour simplifier la représentation, ces éléments de voisinages sont répartis en deux sous-ensembles de mots apparaissant respectivement avant ou après le mot cible dans les contextes étudiés. Il est naturellement possible pour un mot participant à la caractérisation d'être associé à plusieurs séquences de types distinctes, mais si un mot apparaît plusieurs fois associé avec la même séquence de types grammaticaux dans un voisinage de la cible, seule une instance est conservée.

Si nous construisons par exemple la caractérisation du mot « mot », à partir du premier paragraphe de cette section, où il est amplement utilisé, nous obtenons la représentation reproduite sur la figure 3.4. Il est remarquable, à l'étude de cet exemple, que beaucoup des mots retenus entretiennent en effet dans le texte un lien syntaxique avec *mot*, mais que ce ne sont pas forcément ceux-là qui constituent les indices les plus pertinents concernant son sens : ainsi *syntaxique* (qui apparaît dans l'expression « catégorie syntaxique des mots ») ne présente-t-il aucun lien grammatical direct avec *mot*, alors qu'il fournit une indication (au moins domaniale) précieuse sur son sens. Cette constatation nous conforte dans notre choix de considérer des voisinages « plats » d'occurrences de mots plutôt que des contextes syntaxiquement structurés.

Avant		Après	
associer	[préposition]	étudié	[∅]
		ensemble	[adjectif]
étudié	[nom préposition]	apparaître	[∅]
ensemble	[préposition]	voisinage	[verbe préposition]
voisinage	[préposition]	caractériser	[préposition]
catégorie	[adjectif préposition]	susceptible	[∅]
syntaxique	[préposition]	jouer	[adjectif préposition]
fixer	[préposition nombre]		
4	[∅]		
		cible	[∅]
		expérience	[nom préposition]

FIG. 3.4 – *Exemple de représentation « ensembliste typée » des voisinages d'apparition de mot dans un paragraphe de texte*

La similarité entre deux tels « ensembles augmentés » de mots est calculée de manière similaire à précédemment en employant un indice ensembliste de type Jaccard. Cette fois, néanmoins, l'intersection entre deux ensembles est réalisée de manière plus restrictive, puisqu'il est imposé qu'un mot soit présent dans les ensembles caractéristiques des deux mots comparés *et* associé à la même séquence de catégories syntaxiques pour y être inclus. Cette contrainte assez forte a pour but d'augmenter la probabilité que les mots ainsi comptés comme communs à deux représentations aient entretenu le même rapport avec les mots caractérisés. Son intérêt est également de rendre moins critique la question du choix de la taille de voisinage : il existe en effet très peu de chances de voir apparaître dans les représentations de deux mots un voisin identique associé à la même séquence de plus de trois catégories grammaticales — et si cela ce produit, le mot en question n'est pas moins pertinent qu'un autre pour la comparaison. Le critère guidant le choix de la taille de voisinage considéré résulte donc principalement d'un impératif d'efficacité calculatoire, le risque de prendre en compte des indices trop éloignés étant devenu minime.

La technique déployée à la section précédente pour normaliser la matrice de similarité obtenue est également employée pour cette seconde méthode, et les valeurs de similarité ainsi normalisées sont exploitées par le même algorithme de classification ascendante hiérarchique pour la construction d'arbres de classification des mots.

3.4.5.2. Résultats

Le mode de représentation des voisinages décrit dans cette section étant beaucoup plus contraignant que le précédent pour la comparaison de deux

mots, il n'est pas surprenant qu'il présente une efficacité inférieure pour leur rapprochement : ainsi les classes observables dans les arbres de classification produits sont-elles (à seuil de nombre d'occurrences de sélection équivalent) de plus petite taille, plus « éclatées » dans l'arbre que précédemment. On n'observe en revanche que très peu de rapprochements de bas niveau entre mots sans rapport entre eux.

Bien que cette nouvelle méthode ne permette pas d'obtenir des résultats de qualité équivalente à la précédente, elle s'en approche suffisamment pour constituer une alternative viable dans les cas où les données textuelles ne se prêteraient pas au mode de sélection des indices contextuels proposé en 3.4.1. Néanmoins, la qualité des résultats qu'il est possible d'espérer obtenir dans la suite de nos recherches dépendant grandement de la qualité de la proximité entre mots que nous tentons d'évaluer dans cette première étape, nous donnons dans l'immédiat la priorité à cette qualité et retenons comme ressource pour les travaux présentés dans les sections suivantes la mesure de proximité sémantique définie précédemment, à la section 3.4.4.

Ayant proposé une solution satisfaisante pour notre besoin à la première problématique posée d'évaluation de la similarité sémantique entre mots en employant d'importantes quantités de données généralistes, nous nous penchons maintenant sur la seconde question, qui concerne la construction de classes sémantiques à partir de sous-corpus thématiquement homogènes mais de petite taille, en exploitant cette première évaluation de proximité sémantique.

3.5. Classification sémantique des noms sur un sous-corpus thématique

Confronté à la nécessité de définir un mode de représentation des informations contextuelles caractérisant l'usage d'un mot qui puisse être efficace pour l'étude des sous-corpus thématiques que nous considérons désormais, il semble que nous nous attelions à une tâche lourde de contradictions : la représentation désirée doit être à la fois suffisamment rigoureuse pour éviter de collecter de faux indices positifs pouvant mener à rapprocher des mots de manière erronée, et suffisamment souple pour favoriser l'émergence d'indices justifiant des rapprochements entre mots ; elle doit en outre être suffisamment simple dans son principe pour pouvoir exploiter de manière directe la préconnaissance de proximités sémantiques entre mots calculée au chapitre précédent. Or les possibilités de caractérisation synthétique des contextes d'usage d'un mot sont légion, et toutes sensiblement équivalentes les unes aux autres. Toutes présentent en outre une faiblesse théorique importante : rien ne permet de fonder linguistiquement le regroupement d'un ensemble d'observations distinctes dans le texte en une collection centralisée d'indices ; le passage de l'objet linguistique qu'est l'énoncé à l'objet statistique (d'ailleurs plus ou

moins rigoureusement défini de ce point de vue) qu'est la caractérisation globale d'un mot par l'ensemble de ses usages présente de manière systématique une solution de continuité théorique importante.

C'est pourquoi nous avons fait le choix de considérer comme objet d'étude non pas le mot, de caractérisation problématique, mais le voisinage lui-même, beaucoup plus aisé à représenter et dont la comparaison donne bien mieux prise à l'intuition ou à l'analyse linguistique (section 3.5.1). Les similarités entre mots sont ensuite calculées par une technique d'échantillonnage aléatoire que nous avons développée afin de permettre la comparaison non biaisée de mots pouvant présenter des nombres d'occurrences très divers (section 3.5.2). L'exploitation des similarités ainsi calculées par une méthode de classification hiérarchique permet la construction d'arbres de classification d'assez bonne qualité (section 3.5.3), bien qu'une étape de filtrage manuel des résultats soit encore nécessaire pour définir des classes sémantiques à proprement parler.

Nous avons dans la section précédente considéré la classification sémantiques de mots de catégories morphosyntaxiques diverses, guidé en cela par le besoin que nous avions de disposer de premières informations sémantiques pour toutes ces catégories. La construction de classes sémantiques étant dans cette seconde étape une « fin en soi », nous nous limitons par souci de clarté de la présentation à l'étude des seuls noms.

3.5.1. Représentation des voisinages

Les voisinages étant des objets de nature assez élémentaire, peu de choix sont possibles pour leur représentation : la principale décision à prendre consiste à maintenir ou non l'information concernant l'ordre dans lequel les mots apparaissent dans le texte. Nous présentons donc deux représentations possibles des voisinages, l'une séquentielle, l'autre ensembliste.

Représentation séquentielle des voisinages Dans ce cas, le voisinage lui-même est sa propre représentation. Il est néanmoins possible d'écarter de la séquence de mots retenue certaines catégories grammaticales (par exemple les conjonctions et déterminants) ou de substituer à certains mots un simple indicateur de cette catégorie (par exemple en remplaçant toutes les prépositions par `<préposition>`). La comparaison entre séquences de mots doit pouvoir évaluer la proximité sémantique entre celles-ci, ce qui constitue un problème très complexe. Nous avons proposé au cours de nos expériences de l'évaluer par une version spécialisée de la distance d'édition, ou distance de Levenshtein, qui les sépare[4]. La distance d'édition traditionnelle mesure le nombre

[4]Quoique la validité du mode de représentation en lui-même ne dépende pas de la langue considérée, la mesure de similarité que nous proposons pour l'exploiter est, elle, fortement dépendante des structures de la langue considérée.

d'opérations élémentaires (insertion, suppression et substitution) qui sont nécessaires pour transformer une séquence de symboles élémentaires (typiquement, une chaîne de caractères) en une autre. Considérant le mot comme symbole élémentaire, nous avons adapté à nos besoins la distance de Levenshtein par l'ajout des deux règles suivantes :

1. le coût d'une modification est fonction de la catégorie grammaticale du mot manipulé, ces coûts étant définis en fonction de l'évaluation de l'impact sémantique que l'opération peut avoir sur la séquence linguistique (il est par exemple possible de considérer l'insertion ou la suppression d'une préposition comme moins coûteuses que celles d'un verbe),
2. les substitutions ne sont possibles qu'entre mots de même catégorie, et leur coût est proportionnel à la « distance sémantique » (calculée à la section 3.4.4) qui sépare le mot d'origine de son substitut.

Cette seconde règle est une première façon d'introduire de manière assez intuitive et naturelle dans le calcul de la similarité les connaissances acquises lors de l'étape précédente. La normalisation des valeurs calculées par cette méthode pour des séquences de longueurs variables étant assez délicate, nous définissons la similarité entre deux séquences de mots comme la longueur de la séquence commune maximale qu'il est possible d'obtenir pour un coût d'édition arbitraire fixé à l'avance.

Représentation ensembliste des voisinages Cette fois, une sélection de mots composant le voisinage est regroupée sans ordre particulier au sein d'un ensemble. L'information de séquentialité étant perdue lors de cette transformation, les mots tendant à préciser les rapports entretenus par le nom central avec ses voisins (conjonctions, prépositions, *etc.*) ne constituent plus une source d'information pertinente ; en conséquence, la sélection réalisée se limite aux mots « pleins » (noms et noms propres, verbes, adjectifs, adverbes) et aux nombres. La similarité entre deux voisinages est donnée par le simple cardinal de l'intersection de leurs ensembles représentatifs. Le choix de ne pas normaliser cette valeur (par exemple en la divisant par le cardinal de leur union, pour obtenir un indice de Jaccard) s'explique en termes de *contrainte sémantique* exercée sur un mot par son voisinage : V_1 et V_2 étant les ensembles représentant deux contextes, s'ils ne contiennent qu'un mot chacun et ont ce mot en commun, cela reste un indicateur sémantique faible, même si la coïncidence est de 100 %. En revanche, si V_1 et V_2 rassemblent 5 mots chacun, dont 3 « seulement » en commun, ces mots partagés constituent bien un indice de proximité sémantique potentiellement plus fort.

Afin de prendre en compte la connaissance des rapprochements entre mots acquise sur l'ensemble du corpus à l'étape précédente, nous définissons une mesure de similarité entre deux ensembles de mots V_1 et V_2 consistant à chercher pour chacun des éléments d'un ensemble celui qui lui ressemble le plus

dans l'autre. Si p désigne la mesure de proximité entre mots calculée à la section 3.4, nous définissons la similarité ς entre V_1 et V_2 par :

$$\varsigma(V_1, V_2) = \sum_{a \in V_1} \max_{b \in V_2} (p(a, b))$$

Cette mesure présente l'inconvénient d'être asymétrique ; une « symétrisation » par moyenne des valeurs réciproques est donc nécessaire pour obtenir un indice s directement exploitable pour la classification :

$$s(V_1, V_2) = \frac{\varsigma(V_1, V_2) + \varsigma(V_2, V_1)}{2}$$

Les calculs mis en œuvre dans cette section pouvant atteindre une assez grande complexité et travailler sur des quantités de données importantes, la taille de voisinage choisie doit rester relativement restreinte pour que la méthode proposée reste « praticable ». Nous la définissons donc à quatre mots (pleins ou non) à droite et à gauche.

Des deux techniques de représentation présentées, seule la seconde (« ensembliste ») permet d'obtenir des résultats réellement concluants ; c'est elle que nous retenons notamment dans la suite lors de la présentation des classifications sémantiques obtenues. Pour autant, nous souhaitons maintenir dans l'immédiat la possibilité théorique d'utiliser quelque mode de représentation et mesure de similarité entre contextes que ce soit — l'un des intérêt majeurs de la méthode présentée étant sa totale généricité de ce point de vue. C'est pourquoi nous considérons dans les sous-sections suivantes, consacrées à l'évaluation de la similarité entre mots, une mesure de similarité « virtuelle » σ entre leurs voisinages.

3.5.2. Similarité entre mots à partir des similarités entre leurs contextes

Nous définissons dans cette section le mode de calcul de la similarité entre deux mots m_1 et m_2 apparaissant respectivement n_1 et n_2 fois dans le sous-corpus d'étude (autrement dit, caractérisés par les données respectives de n_1 et n_2 voisinages) à partir d'une similarité σ définie entre leurs voisinages. Nous notons $v_{ij,\ 1 \leq j \leq n_i}$ les voisinages du mot m_i.

Une manière « élémentaire » de répondre à cette question consisterait à calculer la moyenne des similarités entre toutes les paires de voisinages de m_1 et m_2 possibles. Néanmoins, cette technique répond mal à la question que nous nous posons, toujours à partir de la définition retenue pour les classes sémantiques : « est-il possible d'utiliser m_2 à la place de m_1 dans les énoncés où ce dernier apparaît, et réciproquement ? ». Pour un voisinage de m_1 donné, la réponse à cette question est positive s'il existe *un* voisinage de m_2 lui ressemblant substantiellement ; il est donc pertinent de prendre en compte pour le calcul uniquement le voisinage de m_2 ressemblant le plus au voisinage de m_1

considéré. Nous calculons donc, pour évaluer le degré de « remplaçabilité » de m_1 par m_2, la moyenne des similarités entre les voisinages de m_1 et ceux qui leur ressemblent le plus parmi les voisinages de m_2 :

$$s\left(m_1, m_2\right) = \frac{\sum_{j=0}^{n_1} \max\left\{\sigma(v_{1j}, v_{2k}) \| 1 \leq k \leq n_2\right\}}{n_1} \tag{3.13}$$

La principale difficulté de cette approche résulte du fait que la mesure ainsi définie n'est pas symétrique : il est possible que de nombreux voisinages de m_1 aient un correspondant parmi ceux de m_2 sans que le contraire soit vrai, formellement :

$$\frac{1}{n_1}\sum_{j=0}^{n_1} \max_k \left\{\sigma\left(v_{1j}, v_{2k}\right)\right\} > \frac{1}{n_2}\sum_{k=0}^{n_2} \max_j \left\{\sigma\left(v_{2k}, v_{1j}\right)\right\}$$

Cette configuration peut s'interpréter, sous réserve de la qualité de la mesure de similarité entre voisinages employée, comme un cas où m_2 peut couramment être employé à la place de m_1, mais pas m_1 à la place de m_2, autrement dit, où m_2 est un hyperonyme de m_1 — mais cela n'est en réalité qu'une supposition théorique, qui se vérifie mal en pratique. Nous n'avons pu mettre au point au cours de notre étude de méthode de classification permettant d'exploiter intelligemment une mesure de similarité asymétrique ; le plus efficace reste de combiner $\sigma(m_1, m_2)$ et $\sigma(m_2, m_1)$ afin d'obtenir une similarité synthétique entre ces deux objets. Empiriquement, les meilleurs résultats sont obtenus en réalisant une moyenne de ces deux valeurs pondérée par le nombre de voisinages sur lequel elles ont été calculées, définissant ainsi la mesure de similarité entre mots s par :

$$s\left(m_1, m_2\right) = \frac{n_1\sigma(m_1, m_2) + n_2\sigma(m_2, m_1)}{n_1 + n_2} \tag{3.14}$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{j=0}^{n_1} \max_k \left\{\sigma\left(v_{1j}, v_{2k}\right)\right\} + \sum_{k=0}^{n_2} \max_j \left\{\sigma\left(v_{2k}, v_{1j}\right)\right\}\right] \tag{3.15}$$

De manière générale, cette mesure est très sensible aux différences de volume de données permettant de caractériser les mots étudiés. La formule choisie limite en particulier très fortement les possibilités de similarité entre m_1 et m_2 si $n_1 \ll n_2$. Afin de contourner cette difficulté, nous n'appliquons la formule donnée par l'équation 3.15 qu'à des mots présentant exactement le même nombre d'occurrences ; cela est rendu possible par une méthode d'évaluation d'une valeur de similarité entre mots par échantillonnage aléatoire de leurs représentations, dont nous présentons maintenant le fonctionnement.

Évaluation par échantillonnage aléatoire de similarités entre mots

La technique présentée ici a pour objectif de permettre le calcul de la similarité entre des objets caractérisés par des quantités de données très inégales. Typiquement, dès que la similarité employée est relativement complexe, il n'est pas possible de trouver un facteur de normalisation simple permettant de comparer « sur un pied d'égalité » les objets étudiés. Nous proposons pour résoudre ce problème de simuler la comparaison d'objets caractérisés par des quantités de données identiques.

Cette technique est inspirée des méthodes d'évaluation par échantillonnage aléatoire comme celle de Monte Carlo [ET91], la principale originalité de notre approche étant de réaliser l'échantillonnage non pas au sein de deux populations d'objets à comparer, mais au sein des indices caractérisant deux objets étudiés.

Soit deux mots m_1 et m_2 à comparer, caractérisés respectivement par la donnée de n_1 et n_2 voisinages. À partir de ces représentations de m_1 et m_2, nous générons une population de k mots « fictifs » (que nous nommons *représentants*) caractérisés chacun par la donnée d'un ensemble de l voisinages tirés au sort (avec remise) parmi les voisinages caractéristiques du mot d'origine. L'intérêt de cette opération est de pouvoir créer des populations de représentants comprenant tous le même nombre de voisinages l ; il est ainsi possible de calculer la similarité entre deux représentants de deux mots distincts en employant directement la formule de l'équation (3.15), sans préoccupation de normalisation.

La similarité entre les deux mots d'origine est définie comme la moyenne des similarités entre toutes les paires possibles de représentants de m_1 et m_2 :

$$S(m_1, m_2) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} s(r_{1i}, r_{2j})}{k^2} \tag{3.16}$$

où r_{1i} et r_{2j} désignent les représentants de m_1 et m_2, respectivement, et k le nombre de représentants générés pour chaque mot.

Dans la pratique, afin d'assurer la représentativité des populations de représentants générées, la procédure de sélection aléatoire des voisinages caractéristiques « force le hasard » afin de s'assurer que chacun des n voisinages d'un mot est présent au moins $\frac{kl}{n}$ fois dans les caractérisations construites pour les représentants de ce mot.

Les valeurs des mesures de similarité calculées par ce procédé se stabilisent lorsque le produit kl devient nettement supérieur aux nombres d'occurrences respectifs des mots comparés. Lors de nos expériences, travaillant sur des mots apparaissant entre 20 et 1000 fois, nous avons adopté les valeurs $k = 200$ et $l = 20$.

La figure 3.5 illustre le bénéfice possible de la méthode proposée dans un

Considérons trois objets $\mathcal{O}_1$, $\mathcal{O}_2$ et $\mathcal{O}_3$ définis comme des ensembles de lettres :

$$\begin{aligned}\mathcal{O}_1 &= \{b, c, l, n, r, t\}\\ \mathcal{O}_2 &= \{a, c, e, g, i, k, m, o, q, s, u, w, y\}\\ \mathcal{O}_3 &= \{a, m, s\}\end{aligned}$$

Afin de réaliser un parallèle avec le type de problème que nous traitons en réalité (comparaison de noms en fonction des similarités existant entre leurs contextes), nous définissons une similarité ς entre lettres, inversement proportionnelle à la différence entre leurs rangs dans l'alphabet, qui forme la base de la similarité entre objets :

$$\varsigma(c_1, c_2) = \frac{1}{1 + |\text{rang}(c_1) - \text{rang}(c_2)|}$$

La représentation ci-dessous des trois objets considérés de manière « déroulée » montre que, quoique que n'ayant aucun élément en commun, $\mathcal{O}_1$ et $\mathcal{O}_3$ semblent plus similaires que $\mathcal{O}_2$ et $\mathcal{O}_3$ — les lettres composant $\mathcal{O}_1$ se trouvent en effet « dans les mêmes zones » que celles définissant $\mathcal{O}_3$, tandis que celles de $\mathcal{O}_2$ ne coïncident avec elles qu'« accidentellement » :

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
$\mathcal{O}_1$		■	■									■		■				■		■						
$\mathcal{O}_3$	■												■						■							
$\mathcal{O}_2$	■		■		■		■		■		■		■		■		■		■		■		■		■	

Pourtant, si nous employons directement la mesure de similarité définie à l'équation 3.15, nous obtenons les valeurs $s(\mathcal{O}_1, \mathcal{O}_3) \simeq 0,48$ et $s(\mathcal{O}_2, \mathcal{O}_3) \simeq 0,53$.
Les deux tableaux présentés ci-dessous montrent le résultat obtenu par la méthode d'échantillonnage aléatoire que nous avons développée. Le choix effectué ici est de générer pour chaque objet 20 « représentants » comprenant chacun deux lettres.

Comparaison $\mathcal{O}_1$ / $\mathcal{O}_3$		
(s, a)	(r, b)	0,50
(m, a)	(l, n)	0,40
(s, s)	(b, n)	0,14
(s, a)	(n, l)	0,14
(s, a)	(t, r)	0,39
(a, s)	(r, c)	0,42
(a, s)	(t, b)	0,50
(s, a)	(b, t)	0,50
(m, s)	(r, r)	0,42
(m, a)	(r, b)	0,33
(a, s)	(n, t)	0,31
(m, m)	(b, l)	0,40
(a, s)	(n, b)	0,33
(s, m)	(l, l)	0,41
(m, m)	(r, b)	0,15
(a, s)	(l, b)	0,31
(a, m)	(n, t)	0,30
(m, a)	(c, l)	0,42
(a, s)	(b, l)	0,31
(s, s)	(r, b)	0,39
(a, s)	(l, r)	0,30
	Moyenne	0,37

Comparaison $\mathcal{O}_2$ / $\mathcal{O}_3$		
(m, a)	(s, q)	0,15
(a, s)	(i, c)	0,22
(m, m)	(c, q)	0,17
(a, s)	(w, e)	0,20
(m, a)	(k, y)	0,21
(m, s)	(k, m)	0,62
(m, m)	(s, k)	0,29
(m, a)	(g, q)	0,17
(a, s)	(q, y)	0,22
(m, a)	(g, s)	0,14
(a, s)	(k, a)	0,56
(s, s)	(k, y)	0,13
(a, s)	(y, k)	0,12
(s, a)	(o, e)	0,20
(s, a)	(g, o)	0,17
(a, s)	(a, w)	0,60
(a, s)	(a, m)	0,57
(s, a)	(e, g)	0,15
(m, m)	(o, a)	0,27
(m, s)	(g, k)	0,23
(m, a)	(y, e)	0,15
	Moyenne	0,28

Le résultat obtenu concorde cette fois avec la perception intuitive des ressemblances.

FIG. 3.5 – *Illustration de l'intérêt du calcul des valeurs de similarité entre objets par échantillonnage aléatoire de leurs caractérisations*

cas concret (quoique artificiel) de comparaison entre objets simples : l'objet $\mathcal{O}_2$ comptant significativement moins d'occurrences que $\mathcal{O}_1$ et $\mathcal{O}_3$, le calcul direct de leur similarité fournit un résultat peu intuitif, que l'approche par échantillonnage aléatoire que nous proposons permet d'amender. Il est sans doute utile de préciser pour finir que la méthode décrite ne présente d'intérêt que dans le cas où la similarité entre les objets à comparer dépend d'une similarité définie sur leurs attributs, comme c'est le cas du problème qui nous concerne et de l'exemple donné. Tenter de l'appliquer pour calculer une similarité entre objets définis, par exemple, par des attributs binaires, ne peut aboutir qu'au même résultat qu'un calcul direct (aux imprécisions introduites par les imperfections d'échantillonnage près).

3.5.3. Résultats

Appliquée à une sélection des 300 noms les plus fréquents du domaine des « nouvelles technologies » (noms apparaissant plus de 20 fois dans les données considérées, le plus fréquent apparaissant quelque 2 000 fois), la méthode de comparaison entre noms définie dans cette section permet la construction de l'arbre dont la figure 3.6 présente la structure générale. Si nous évaluons sa qualité de la même manière que nous l'avons fait en section 3.4.4 pour la méthode appliquée à la totalité des données du corpus, le résultat en est bien moins probant : environ un tiers des mots classés s'organisent en classes pertinente si la tâche de lecture manuelle se limite à la sélection des meilleurs nœuds. Il est nécessaire, si nous souhaitons constituer des classes sémantiques réellement intéressantes et pouvant constituer une proposition de structuration d'un lexique sémantique, d'étendre l'intervention humaine au-delà de ce simple rôle de sélection, et de permettre un léger filtrage manuel écartant quelques mots parasites empêchant la formation de « bonnes » classes. Sous ces conditions, il est possible de rendre manifestes les classes que nous avons fait apparaître en noir sur la figure 3.6 — les parties en gris correspondant aux « mots parasites » écartés manuellement ainsi qu'aux « mauvaises » classes. Les classes valides ainsi extraites rassemblent alors environ 60 % des mots classés.

Il est malheureusement impossible, pour des raisons calculatoires, d'appliquer la méthode présentée dans cette section à la totalité des données du corpus du *Monde diplomatique* afin d'évaluer son apport dans ce cas. Il serait en effet intéressant d'évaluer si la technique que nous avons développée pour travailler sur un petit sous-corpus thématique, et qui dans ce cadre fournit un résultat très nettement supérieur à la méthode « générale » développée pour le corpus complet, dépasse également cette dernière si on l'applique au corpus intégral, ou si ces deux techniques sont chacunes particulièrement adaptées à un type de corpus particulier (multi- ou monothématique) et ne fonctionnent efficacement que sur celui-là.

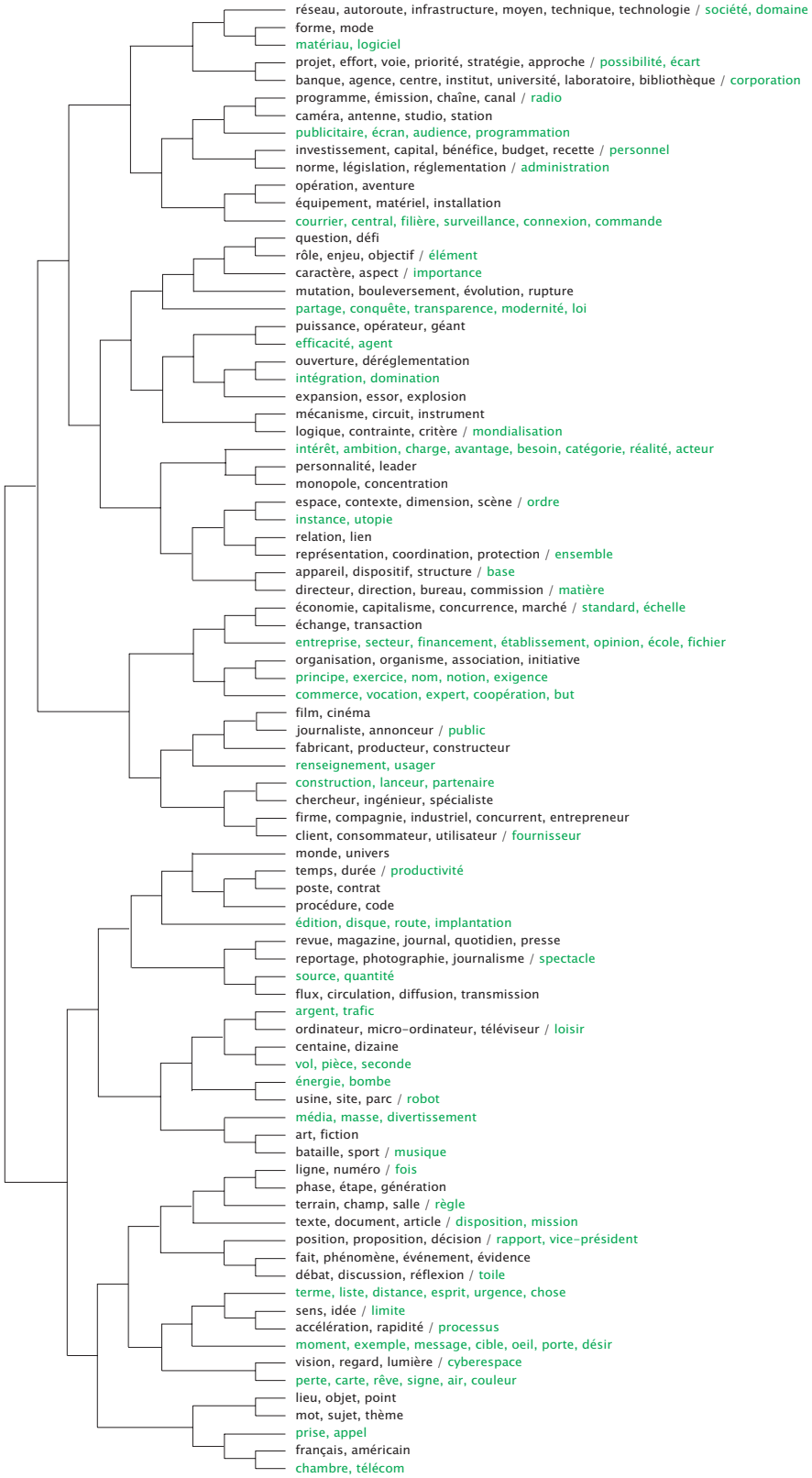


FIG. 3.6 – *Structure générale de l'arbre de classification sémantique des 300 noms les plus fréquents du domaine « nouvelles technologies », obtenu par analyse du sous-corpus associé employant la méthode décrite à la section 3.5. Nous avons extrait manuellement de l'arbre les classes qui nous semblaient les plus pertinentes (en noir), laissant de côté les autres (en gris)*

3.6. Conclusion

Nous avons présenté dans ce chapitre une méthodologie générale pour l'acquisition automatique, à partir de sous-corpus de langue non spécialisée mais thématiquement homogènes, de classes sémantiques structurant le domaine lexical associé au thème abordé dans les données textuelles employées. La méthode proposée est composée de deux étapes :

– la première consiste à réaliser, en employant la totalité du corpus général, un apprentissage de proximités sémantiques entre mots, par comparaison des ensembles de mots apparaissant dans les voisinages de leurs occurrences ;
– la seconde fait usage de cette connaissance afin de mesurer les similarités entre les voisinages des occurrences de mots à classer, puis extrapole à partir de ces similarités entre voisinages des similarités entre mots permettant une classification hiérarchique de ceux-ci.

Cette structure originale nous permet de mettre au jour à partir de corpus de tailles assez restreintes des classes sémantiques réellement pertinentes et spécifiques à un domaine : on peut par exemple noter sur la figure 3.6 le rapprochement de *réseau* et *autoroute*, révélateur du sens particulier donné à *autoroute* dans le domaine « nouvelles technologies » (« autoroutes de l'information »).

Nos recherches ont en outre occasionné le développement de deux méthodes statistiques originales. La première permet la normalisation *a posteriori* de valeurs de similarité rassemblées dans une matrice, et s'avère un outil précieux pour l'exploitation de formules de calcul de similarité difficilement normalisables *a priori*. La seconde, définie afin de calculer des similarités entre mots à partir des similarités existant entre leurs contextes, constitue selon nous une contribution importante au domaine de l'acquisition automatique sur corpus de classes sémantiques. Elle permet en effet de réaliser une segmentation nette entre le travail proprement linguistique de comparaison de contextes d'apparition de mots, et le processus statistique de généralisation de ces comparaisons à l'échelle de l'ensemble des contextes qui caractérisent l'usage d'un mot. Nous n'avons eu le loisir de développer au cours de nos recherches que deux types de représentation des voisinages d'occurrences de mots, ainsi que des mesures de similarité relativement simples associées à ceux-ci, mais il ne s'agit que d'une introduction aux développements que le cadre méthodologique proposé rend possibles. Il est désormais nécessaire, afin de mener ces premiers travaux au maximum de leur potentiel, de réaliser une étude linguistique rigoureuse des critères pouvant être retenus pour affirmer que deux voisinages imposent à leur mot central des contraintes sémantiques similaires.

Une question importante se pose à l'issue de cette étape de nos recherches : les classes sémantiques que nous avons constituées correspondent-elles réellement aux taxèmes définis par la sémantique différentielle ? Il n'est possible

de répondre que partiellement à cette interrogation. Sur le plan des certitudes, tous les calculs mis en œuvre dans ce chapitre tendent à rapprocher des mots employés de manière similaire ; or, le test de permutabilité en contexte constitue le fondement même de la définition d'un taxème : il n'y a donc aucune raison de soupçonner que les mesures de similarité définies puissent regrouper des mots de sens proche selon une optique autre que taxémique[5]. En revanche, nous ne disposons pas de critère qui nous permettrait de définir quelle est la « granularité de sens » optimale pour constituer un taxème — et F. Rastier reste lui-même dans ses ouvrages assez vague à ce sujet. Dans l'exemple de l'arbre de classification de la figure 3.6, est-il réellement pertinent de regrouper comme nous le faisons *investissement*, *capital*, *budget*, *bénéfice* et *recette*, ou serait-il plus judicieux de distinguer les sommes d'argents « fixes » {*capital*, *budget*, *investissement*} des « rentrées » {*bénéfice*, *recette*} ? Rien ne permet *a priori* d'en juger.

La seule réponse totalement satisfaisante à la question du choix d'un degré de généralité pour les taxèmes consisterait à développer un mode de lecture automatisée des arbres de classification ne s'appuyant que sur les données textuelles elles-même, à l'instar de ce qui a été fait pour le système FAESTOS. Nous n'avons pu au cours de nos travaux mettre au point un tel outil, et en sommes réduit à définir une « ligne directrice » pour systématiser l'exploitation manuelle des arbres de classification produits : sélectionner les plus grandes classes possibles telles qu'un élément de sens « évident » justifie le regroupement, s'en remettant au pouvoir discriminant des sèmes spécifiques pour structurer les ensembles ainsi constitués.

Nous considérons malgré ces limitations et réserves les classes sémantiques extraites manuellement de l'arbre de classification construit automatiquement grâce aux outils présentés dans ce chapitre comme une première structuration des domaines suffisamment pertinente pour constituer une approximation satisfaisante de taxèmes — et employons dorénavant ce nom pour les désigner. Nous entreprenons au chapitre suivant de rechercher en leur sein des sèmes spécifiques permettant de distinguer plus finement les sens des mots rassemblés.

[5] Optique qui resterait à définir.

Chapitre 4

Structuration de taxèmes par des sèmes spécifiques

We penetrated deeper and deeper into the heart of darkness. It was very quiet there.

J. Conrad, *Heart of Darkness* [Con02]

4.1. Introduction

Contrairement aux travaux présentés aux chapitres précédents, qui se rattachent à des domaines de recherche déjà actifs, les expériences relatées à ce chapitre abordent une question encore très peu étudiée : celle de la recherche au sein d'un ensemble de mots de sens proches de distinctions précisant les nuances qui les individualisent. Bien que la structuration d'un vocabulaire en taxèmes puisse constituer un outil précieux pour de nombreuses tâches, on ne peut en effet que constater son insuffisance si l'objectif poursuivi requiert une interprétation assez précise du texte : *technologie* et *technique*, tout comme *moderne* et *ancestral* sont sans doute théoriquement échangeables en contexte, mais approximer « technologie moderne » par « technique ancestrale » en recherche d'information, par exemple, risque d'aboutir à des résultats peu satisfaisants pour l'utilisateur. Similairement, B. Habert *et al.* notent dans [HNN96] que, si *coronaire* et *coronarien* sont d'après le système ZELLIG les adjectifs les plus similaires du corpus médical qu'ils étudient — proximité confirmée par une compréhension « de bon sens » —, les connexions de ces mots dans le graphe (seul *coronarien* est lié à *sévère*, *significatif* et *important*) suggèrent qu'il n'est pas toujours possible d'employer l'un pour l'autre. Il semble donc nécessaire, afin de rendre les taxèmes réellement utilisables, d'en réaliser une structuration interne susceptible de rendre compte des restrictions de sélection additionnelles qui encadrent plus rigoureusement l'usage des mots — autrement dit, des nuances qui précisent leurs sens.

Les recherches entreprises dans ce sens s'intéressent, à notre connaissance, exclusivement à la mise en évidence de relations lexicales sémantiques « classiques » entre mots, tentant en particulier d'identifier les « vrais » synonymes ainsi que les couples en relation d'hyperonymie / hyponymie. Pourtant, la synonymie authentique, c'est-à-dire la possibilité de substitution *systématique* d'un mot par un autre, est un phénomène très rare en pratique (principalement limité aux variantes d'écriture, comme dans l'alternative *cd-rom* / *cédérom*). L'hyperonymie, pour sa part, ne constitue pas un indice permettant de juger de la pertinence d'un usage de mot en contexte : il est par exemple rarement heureux, même si théoriquement possible, de remplacer dans un texte *chien* par *animal*. Le type de distinctions sémantiques que nous souhaitons mettre au jour est à la fois plus subtil et plus « pratique » que ces relations ; l'objectif est de pouvoir préciser, par exemple, que *technologie* se distingue de *technique* par une sophistication supérieure (comme l'atteste la définition proposée par le *Petit Robert* [LRDR93] : « technique moderne et complexe »), ou que *coronarien* s'associe plus volontiers à la désignation d'une maladie ou d'un symptôme, et *coronaire* à celle d'un organe.

Le formalisme de représentation du sens proposé par la sémantique différentielle permet de codifier de telles distinctions grâce aux *sèmes spécifiques*. Ceux-ci sont définis comme étant les relations entre mots précisant au sein d'un taxème les particularités sémantiques qui justifient qu'en contexte, alors que tous les sémèmes sont en théorie interchangeables, l'un est choisi de préférence à tous les autres. D'après cette définition, un lexique structuré suivant les principes de la sémantique différentielle doit pouvoir rendre compte des phénomènes interprétatifs de manière beaucoup plus précise que les thesaurus hiérarchiques qui constituent le modèle le plus répandu de représentation de l'information sémantique lexicale exploitable par des outils de TAL. Nous nous intéressons donc dans ce chapitre à la dernière des trois étapes de construction d'un tel lexique : la détermination à partir des données textuelles des sèmes spécifiques distinguant les sens de mots au sein des taxèmes constitués au chapitre précédent. De même que dans la seconde partie du chapitre précédent, nous limitons ici notre étude au seul cas des noms —, sur lesquels s'est concentrée la majorité de nos recherches.

Nous présentons ici dans un premier temps (section 4.2) plusieurs travaux consacrés à la recherche de relations sémantiques lexicales classiques entre mots au sein de classes prédéfinies, qui se trouvent être les plus proches de nos préoccupations. La section 4.3 est consacrée à un retour aux principes théoriques de la sémantique différentielle, ce qui nous permet de préciser les diverses approches possibles de la problématique traitée. Nous présentons ensuite à la section 4.4 un travail précurseur visant à mettre au jour des sèmes spécifiques à partir des voisinages immédiats des occurrences de mots [PS99] ainsi que les expériences que nous avons réalisées afin de tenter d'en approfondir le potentiel. Les maigres résultats que celles-ci permettent d'atteindre de manière automatique nous amènent à identifier plusieurs obstacles théoriques

à cette approche et à en proposer une nouvelle faisant au contraire usage de contextes larges, favorisant l'observation d'isotopies sémantiques « à longue portée » (section 4.5). Cette méthode requiert un travail manuel de sélection des résultats proposés, mais permet à ce prix une structuration cohérente d'un ensemble de taxèmes par des sèmes spécifiques. Nous présentons donc une représentation générale des informations lexicales acquises concernant le domaine « nouvelles technologies », puis une évaluation plus rapide des résultats observés pour d'autres domaines (section 4.6). Nous concluons enfin par un retour sur les enseignements de ces premiers résultats pour la mise au jour de sèmes spécifiques, en esquissant les pistes de recherches les plus prometteuses.

4.2. Relations lexicales « classiques »

Nous regroupons sous la dénomination de relations lexicales sémantiques « classiques » les relations simples et « normalisées » telles que celles couramment rencontrées dans les thesaurus : la synonymie est la plus courante, et les liens d'hyperonymie / hyponymie sont également fréquemment employés. Moins fréquente, la méronymie (relation de la partie au tout) est par exemple codifiée, quoique de manière lacunaire, dans la base de données lexicale WordNet. Cette section est consacrée à la présentation de travaux visant explicitement la recherche de telles relations au sein de classes sémantiques préexistantes.

La préoccupation de structuration de classes sémantiques par des relations lexicales classiques apparaît pour la première fois (à notre connaissance) dans [Gre93], où G. Grefenstette structure en trois parties les listes de termes associées par analogie d'usage à un mot : selon un principe simple exposé dans [FBY92], les mots nettement plus fréquents que le mot-cible dans le corpus étudié sont considérés comme plus généraux (« hyperonymes »), les moins fréquents comme « hyponymes », et les mots de fréquence proche sont placés sur un même niveau de généralité, sans que cela implique pour autant une proximité sémantique plus forte. Sur un corpus médical, par exemple, les mots rapprochés de *cancer* se rassemblent en trois groupes : *maladie* et *tissu* sont plus fréquents, *carcinome* moins fréquent, et *lésion* et *tumeur* présentent un nombre d'occurrences proche.

Ce principe de prise en compte de la fréquence d'un mot pour évaluer sa généralité n'est pas par la suite approfondi par l'auteur, mais est repris par A. Caraballo et E. Charniak dans [CC99], article consacré à l'évaluation de la spécificité de sens de noms. Les auteurs réalisent une comparaison méthodique des indices statistiques employables pour mesurer le degré de généricité des noms d'un corpus à partir de leurs usages. Outre la fréquence des noms, sont considérés le nombre de modifieurs avec lesquels ils sont employés, l'entropie des distributions de ces modifieurs (l'idée étant que plus un nom a un sens général, plus ce sens est susceptible d'être précisé en contexte), ainsi que

l'entropie des distributions des mots apparaissant dans les voisinages de leurs occurrences, la taille de ces voisinages variant de 5 à 50 mots (l'idée étant cette fois qu'un mot générique peut être employé dans une grande quantité de situations). Ces mesures sont appliquées à la structuration hyperonymique de classes sémantiques, chacune des classes étudiées étant construite en extrayant du thesaurus WordNet l'ensemble des hyponymes d'un terme donné ; la hiérarchie proposée par WordNet sert naturellement de référence pour l'évaluation des performances des mesures proposées. Les résultats obtenus montrent qu'une simple mesure de fréquence permet de désigner dans plus de 80 % des cas le plus général de deux mots, cette performance étant améliorée de quelques points par la mesure d'entropie des distributions des modifieurs (qui est en moyenne la plus efficace) et celle de l'entropie des mots des voisinages pour une taille de voisinage importante (50 mots).

Abordant une problématique plus large que ce travail, deux stages menés à l'IRISA en parallèle à nos travaux étudient la structuration interne de classes sémantiques par classification hiérarchique des mots qu'elles regroupent. L'objectif est de faire ainsi apparaître, d'une part, les « vrais synonymes » (ou du moins les mots les plus proches de l'être dans la classe), et d'autre part ceux pouvant jouer le rôle d'hyperonyme par rapport aux autres — ces derniers pouvant également constituer une désignation générale de la classe, à la limite un nommage du sème générique qui la fait naître.

Y.-P. Denielou explore dans [Den04] les possibilités de classifications hiérarchiques à l'intérieur d'une classe sémantique en exploitant des données de voisinage similaires à celles dont nous avons fait usage pour la construction des taxèmes (chaque mot à classer est caractérisé par l'ensemble des mots apparaissant à proximité d'une de ses occurrences dans le texte). À l'issue d'expériences conduites en employant plusieurs types d'indices de similarité, l'auteur est amené à introduire une méthode de classification hiérarchique originale consistant à regrouper à chaque étape non pas les deux objets les plus similaires, mais ceux dont la similarité moyenne aux autres objets à classer est la plus faible — cela de manière cohérente avec l'objectif sous-jacent de recherche de distinctions entre mots. Combiné avec une similarité inspirée de l'indice de Manhattan* appliqué aux vecteurs de fréquences relatives des mots indices dans les contextes des mots classés, cette technique permet par exemple de faire apparaître pour la classe sémantique { *évolution, progrès, mutation, révolution, explosion, innovation, bouleversement* }[1] la hiérarchie présentée à la figure 4.1.

On voit apparaître dans l'arbre de classification ainsi construit des rapprochements et distinctions qui « font sens » du point de vue de l'usage des mots : les paires { *progrès, innovation* } et { *bouleversement, révolution* } obtenues au plus bas niveau de l'arbre correspondent bien aux mots présentant les sens les

[1] Classe construite sur le sous-corpus consacré aux « nouvelles technologies » grâce aux méthodes présentées au chapitre 3.

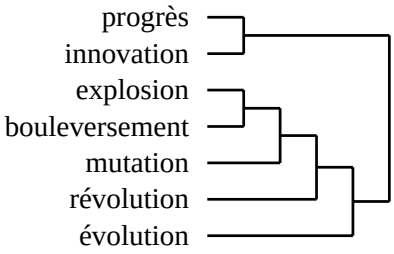


FIG. 4.1 – *Structuration de la classe sémantique { évolution, progrès, mutation, révolution, explosion, innovation, bouleversement } par la méthode de classification hiérarchique proposée dans [Den04]*

plus proches. Même si cet aspect n'est pas explicitement évoqué dans [Den04], la hiérarchie permet ainsi d'isoler les sémèmes distingués des autres mots de la classe par ce qui s'apparente à des sèmes spécifiques : / scientifique / pour la première paire, et / brutal / pour la seconde. En outre, *évolution*, candidat le plus légitime au rôle d'hyperonyme pour l'ensemble de la classe (à défaut de présence dans celle-ci d'un mot réellement général, par exemple *changement*), est le dernier individu agrégé : selon la méthode de classification employée (qui rassemble à chaque étape les mots les moins similaires au reste de la classe), cela suggère qu'*évolution* présente avec tous les mots de la classe une similarité moyenne, indice d'un sens assez général.

Ces premiers résultats intéressants souffrent toutefois d'un manque de « répétabilité » : appliquée à d'autres classes sémantiques, la technique proposée fournit des résultats de qualité assez variable, allant du très bon au totalement inintelligible. M. Coutaud [Cou05] s'appuie sur ces premières expériences et, travaillant sur le même matériau (une sélection de classes sémantiques proposées par le système décrit au chapitre précédent), propose une évolution permettant d'améliorer leurs résultats. La méthode décrite caractérise chaque mot de la classe sémantique à structurer par l'ensemble des adjectifs apparaissant dans une fenêtre de deux mots à droite et à gauche de ses occurrences. Ces données sont exploitées par une méthode mixte combinant classification hiérarchique (dans une version plus classique que celle développée dans [Den04]) et analyse factorielle des correspondances (AFC), grâce à laquelle sont construits des arbres de classification structurant les classes sémantique étudiées de manière plus systématiquement satisfaisante.

La figure 4.2 présente le résultat obtenu pour la même classe { *évolution, progrès, mutation, révolution, explosion, innovation, bouleversement* } ; la structuration est similaire, mais la paire { *bouleversement, explosion* } est enrichie du mot *révolution*, et *évolution* se trouve cette fois clairement regroupé en dernier, au plus proche de la racine de l'arbre. Sur la classe { *échange, circulation, flux, transmission, transport* } (voir figure 4.3), *flux* et *transmission* (qui concernent plus particulièrement des données) se distinguent d'*échange* et *cir-*

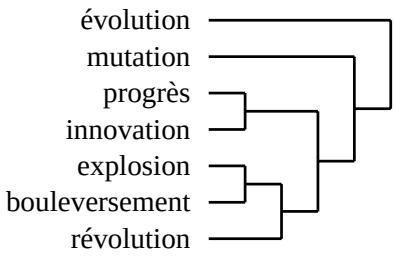


FIG. 4.2 – *Structuration de la classe sémantique { évolution, progrès, mutation, révolution, explosion, innovation, bouleversement } par la méthode mixte AFC / classification hiérarchique proposée dans [Cou05].*

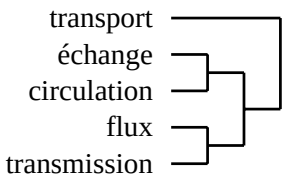


FIG. 4.3 – *Structuration de la classe sémantique { échange, circulation, flux, transmission, transport } par la méthode mixte AFC / classification hiérarchique proposée dans [Cou05].*

culation (qui s'appliquent plus volontiers à des marchandises). Le placement de *transport* en position d'hyperonyme est plus discutable, mais se justifie en l'absence d'un candidat plus naturel (par exemple *déplacement*). La qualité de ces résultats est néanmoins tempérée par leur grande sensibilité aux conditions expérimentales, en particulier les modifications de taille de voisinage ou de mode de sélection des indices. De plus, dans 4 des 5 cas de classification présentés dans [Cou05], le nom placé en « tête » d'arbre de classification, c'est-à-dire en position d'hyperonyme, est également le plus fréquent sur l'extrait étudié (la seule exception étant le cas discutable, déjà évoqué, de *transport*) ; ce résultat n'est pas critiquable en soi, puisque la désignation du mot le plus général est bien pertinente, mais il rend difficile l'évaluation de l'apport réel de la méthode proposée par rapport aux travaux présentés auparavant.

De notre point de vue « sémantique différentialiste », l'intérêt des travaux présentés dans [Den04] et [Cou05] se situe principalement dans les rapprochements qu'ils réalisent entre mots : on a pu constater qu'ils faisaient apparaître dans la classe des « évolutions » des traits de spécificité domaniale (*progrès* et *innovation* s'emploient en particulier dans un contexte scientifique) ou de nature (*bouleversement*, *explosion*, *révolution* : brutalité). Nous souhaitons néanmoins nous écarter de l'approche « classificatrice » qu'ils adoptent pour une raison d'ordre méthodologique relevant de notre volonté de rester aussi fidèle que possible aux principes théoriques de la sémantique différentielle : selon ceux-ci, en effet, les sèmes spécifiques n'ont pas vocation à générer des

sous-classes au sein des taxèmes, mais à marquer les distinctions entre paires de sémèmes. Les exemples présentés dans [Den04] et [Cou05] suggèrent une interprétabilité en termes de sèmes spécifiques des regroupements réalisés, mais il est certain que les méthodes décrites ne peuvent proposer de solution acceptable si les distinctions recherchées correspondent à des relations antisymétriques (par exemple dans le cas d'une gradation de taille). Ainsi, dans l'exemple du taxème // cours d'eau // présenté au chapitre 1, il est impossible de faire apparaître en employant le type de classification proposé par ces travaux la distinction / grand / : en effet, faudrait-il pour cela regrouper prioritairement *ruisseau* et *rivière* (qui sont tous deux « petits par rapport à quelque chose »), ou « rivière » et « fleuve » (« grands par rapport à quelque chose ») ?

Nous attachant à une définition purement différentielle des sèmes spécifiques, nous sommes confronté à la difficulté de devoir représenter une entité linguistique qui n'existe par cette définition qu'« entre » mots, et n'est donc pas exprimée explicitement. La section suivante présente les réponses envisageables à cette question.

4.3. Représentation des sèmes spécifiques

Nous souhaitons dans cette section étudier les méthodes utilisables pour représenter des sèmes spécifiques, en faisant le moins de présupposés possible concernant leur nature. Nous revenons pour cela à leur définition élémentaire comme relations binaires entre sémèmes d'un même taxème, marquant les distinctions sémantiques fines qui justifient qu'en contexte on fasse usage de l'un de ces termes de préférence aux autres[2]. Il résulte d'une interprétation directe de cette définition que les sèmes spécifiques ne peuvent être désignés que de manière implicite, par la donnée des paires de sémèmes qu'ils distinguent. Cette première représentation possible, dont nous approfondissons dans un premier temps le principe (section 4.3.1), est rigoureuse mais peu « intuitive » : si les ressources lexicales sémantiques que nous tentons d'acquérir ont pour vocation d'être employées par un utilisateur humain (en aide à l'interprétation, par exemple), il semble préférable de pouvoir les « étiqueter » d'une manière qui fasse sens. La section 4.3.2 présente les possibilités d'extension dans cette direction autorisées par la sémantique interprétative. Ces deux familles de représentation des sèmes spécifiques suggèrent deux approches assez radicalement différentes de leur mise au jour, comme nous le précisons à la section 4.3.3.

[2]Cette formulation peut laisser supposer qu'un unique sème spécifique distingue chaque élément d'un taxème de tous les autres, ce qui n'est pas le cas. Il est possible et même courant que plusieurs sèmes spécifiques participent simultanément au processus de sélection du « mot juste » en contexte.

4.3.1. Représentation implicite

Toute relation mathématique $\mathcal{R}$ définie sur un ensemble $\mathcal{E}$ peut être représentée par l'ensemble des paires ordonnées $(a, b) \mid \{a, b\} \in \mathcal{E}^2$, telles que $a \mathcal{R} b$. Dans le cas particulier où $\mathcal{R}$ est une relation d'équivalence (symétrique, transitive et réflexive), elle définit une classe d'équivalence : c'est le cas des sèmes génériques, que nous avons mis au jour au chapitre précédent en constituant les taxèmes qui leur sont associés. Dans ce cas, la représentation ensembliste est relativement « compacte » et facilement interprétable.

Dans le cas des sèmes spécifiques, en revanche, il n'est pas possible en théorie (dans une approche purement mathématique) de simplifier la représentation. Il semble donc qu'ils ne puissent être désignés que par l'ensemble des paires de mots qu'ils permettent de distinguer : ainsi l'ensemble de paires { (*rivière*, *ruisseau*), (*montagne*, *colline*), (*immeuble*, *maison*) }, en rassemblant des paires de noms tous distingués l'un de l'autre par / grand /, constitue-t-il une représentation possible de ce sème spécifique. Nous nommons cette représentation « implicite », car elle suppose un travail humain d'interprétation, voire de compréhension, afin de verbaliser le sème — dans l'exemple donné, ni « grand » ni aucun mot de sens proche n'est directement présent dans la représentation. Les relations linguistiques sont néanmoins plus « souples » que les relations mathématiques, et il est courant qu'un sème jouant relativement à un certain contexte de définition un rôle spécifique puisse être générique dans un autre. Cette propriété, que nous précisons ci-dessous, nous permet d'espérer pouvoir donner aux sèmes spécifiques une désignation explicite[3].

4.3.2. Représentation explicite

Les travaux présentés à la section 4.2 montrent qu'il est envisageable d'extraire des taxèmes un terme de plus grande généralité (hyperonyme plus ou moins « parfait » des autres mots), qui peut être considéré comme une dénomination permettant de les désigner de manière intuitive et non ambiguë : le taxème { *évolution*, *progrès*, *mutation*, *révolution*, *explosion*, *innovation*, *bouleversement* } pourrait ainsi être appelé « classe des évolutions ». Un tel processus de nommage, qui nous permettrait idéalement de découvrir pour le sème spécifique donné en exemple à la sous-section précédente l'expression « grand », est également envisageable dans le cas des sèmes spécifiques, mais de manière moins directe.

L'exemple donné au chapitre 1 (figure 1.2) de la structuration du taxème // cours d'eau // par des sèmes spécifiques permettait d'en identifier deux types : les relations antisymétriques comme / grand /, n'ayant de sens que relativement à leurs deux « extrémités », et ceux qui comme / tumultueux / peuvent à la limite

[3] On peut remarquer que la distinction proposée entre représentations implicite et explicite des sèmes rejoint la différence en mathématiques entre réprésentations « en extension » et « en intention » des relations.

être vus comme un attribut du sémème qu'ils distinguent des autres, mais sont de trop faible généralité pour faire naître un taxème indépendant. Nous laissions alors entendre, faute de précision supplémentaire, que / grand / ou / tumultueux / étaient « intrinsèquement » spécifiques, ce qui n'est à l'évidence pas le cas si nous changeons de référentiel :

- considérant le taxème d'adjectifs { *haut*, *grand*, *énorme*, *large*, *gros* }, / grand / est bien sûr le sème générique sous-tendant ce regroupement ;
- le cas de / tumultueux / est plus délicat à illustrer, mais l'on peut par exemple penser aux descriptifs de parcs de loisirs, qui rassemblent couramment les attractions en fonction de la quantité de secousses qui attend leur utilisateur : / tumultueux / pourrait ainsi sous-tendre la classe des manèges les plus « acrobatiques ».

Pour reprendre la formulation de F. Rastier ([Ras96], p.52) : « l'opposition spécifique / générique est *relative* à son ensemble de définition ; aucun sème n'est donc par nature spécifique ou générique ». En particulier, de nombreuses isotopies présentes dans les textes seront dites *mixtes*, le sème les générant assumant dans les mots qui le portent un rôle parfois générique, parfois spécifique. Il est donc possible d'espérer mettre au jour par une analyse des voisinages des mots étudiés de telles isotopies, au sein desquelles pourront être découvertes des expressions dénotant explicitement leurs sèmes spécifiques. Il n'est pas nécessaire en première approximation d'aller pour cela jusqu'à la recherche des taxèmes complets définissant ces sèmes « en tant que génériques » : un unique mot (« grand », voire « gros », dans l'exemple de la sous-section précédente) constituerait d'ores et déjà une représentation concrète lisible et exploitable d'un sème spécifique.

4.3.3. Lien entre mode de représentation et méthodologie de recherche

Du point de vue des techniques d'analyse de données textuelles pouvant permettre la mise au jour de sèmes spécifiques, ces deux grandes familles de représentation — pour lesquelles nous retenons dans la suite les dénominations « implicite » et « explicite » — suggèrent des approches très différentes.

La construction d'une représentation implicite de sèmes spécifiques suppose de regrouper des paires de mots dont les éléments respectifs voient leurs sens distingués par les mêmes différences d'usage. L'optique est donc dans ce cas « classificatrice », les objets étudiés (paires de mots) relevant d'un niveau de complexité supérieur aux simples mots qui faisaient l'objet de l'analyse statistique dans les chapitres précédents. Malgré ce changement de nature des objets, les méthodes employées restent les mêmes : un calcul de similarités entre objets est suivi d'une recherche des similarités maximales. Une conséquence de cela est qu'il est nécessaire, pour que l'étude soit représentative, qu'un nombre assez important de paires de mots (et donc de taxèmes, au sein dequels ces paires soivent être distinguées) soit considéré simultanément.

La mise au jour d'éléments de représentation explicite des sèmes spécifiques, en revanche, implique de ne plus considérer comme objets d'étude les mots du taxème caractérisés par leurs contextes d'usage, mais d'explorer ces contextes eux-mêmes à la recherche d'indices « directs ». Contrairement au cas de la représentation implicite, il est *a priori* tout à fait possible de réaliser cette étude en ne considérant qu'un taxème, ou même une paire de mots isolée.

C'est l'approche explicite qui a été adoptée par les auteurs de [PS99], dont le travail consiste à étudier l'ensemble des éléments de contexte proches des occurrences de chacun des noms d'un taxème (en particulier les adjectifs) afin de tenter d'en extraire une caractérisation distinctive de leurs sens. Nous présentons à la section suivante les résultats de cette recherche et de plusieurs expériences que nous avons menées sur ses traces.

4.4. Recherche de sèmes spécifiques dans les voisinages immédiats

Les expériences que nous relatons dans cette section sont inspirées de la méthodologie décrite par R. Pichon et P. Sébillot dans [PS99], seul recherche existante ayant à notre connaissance abordé la question de la structuration automatique de classes sémantiques par des distinctions du type « sème spécifique ». Les techniques mises en œuvre suivent une approche « traditionnelle » en acquisition d'informations sémantiques lexicales à partir de corpus, consistant à caractériser les mots étudiés par leurs voisinages immédiats dans les textes considérés.

Nous rappelons tout d'abord dans cette section le travail de [PS99], avant de décrire les diverses expérimentations que nous avons réalisées à sa suite, notamment dans l'objectif d'automatiser autant que possible le processus d'analyse et d'interprétation des indices recueillis. Nos recherches dans ce sens ayant abouti à des résultats peu satisfaisants, nous revenons à la section 4.4.4 au guidage par la théorie linguistique : les principes de celle-ci nous permettent en effet d'expliquer les difficultés d'automatisation des méthodes de recherche directe de caractérisations explicites de sèmes spécifiques faisant usage de voisinages immédiats.

4.4.1. [PS99] : une première expérience de structuration de taxème par des sèmes spécifiques

Les auteurs de [PS99] abordent une problématique plus large que celle qui nous préoccupe dans ce chapitre : ils s'intéressent en effet, d'une part, à la détermination des différentes facettes du sens d'un mot suivant le thème du segment de texte où il apparaît, d'autre part, à la recherche d'une structuration sémique d'un taxème explicitant le sème générique qui le sous-tend ainsi que les sèmes spécifiques qui précisent les sens de mots en son sein. La première

de ces deux études montre qu'il est possible, en combinant analyse thématique et exploration des voisinages d'occurrences de mots, d'automatiser partiellement l'identification des différentes facettes de sens d'un même mot telles que celles découvertes par C. Fabre *et al.* dans [FHL97]. Nous ne nous étendons pas ici sur la description de ces expériences, qui appartiennent à une direction de recherche pour ainsi dire « perpendiculaire » à la nôtre et ayant depuis donné lieu à des travaux plus approfondis [Fol02], pour nous concentrer sur la seconde problématique, celle de structuration de taxèmes.

Comme nous, les auteurs de [PS99] réalisent la construction de classes sémantiques au sein de domaines, en exploitant des sous-corpus thématiques associés à ceux-ci. Afin de mettre au jour le sème générique d'un taxème ainsi que les sèmes spécifiques précisant les sens des mots qu'il rassemble, ils caractérisent chacun de ces mots par la liste des noms et adjectifs apparaissant dans une fenêtre de 5 mots avant et après leurs occurrences dans le sous-corpus thématique d'étude. Ces listes sont ensuite écourtées afin de ne retenir pour chaque mot étudié que les 20 mots-indices apparaissant le plus fréquemment dans ses voisinages. La recherche des sèmes est ensuite effectuée en interprétant des sous-listes de voisins extraites par diverses opérations ensemblistes : l'intersection de tous les ensembles d'indices des mots d'une classe permet d'identifier un « tronc commun » de voisins d'où est extrapolé le sème générique, tandis que des différences ensemblistes permettent d'identifier pour chacun des sémèmes considérés ses voisins propres et distinctifs. C'est par comparaison et interprétation manuelles de ces listes que sont découverts les sèmes spécifiques recherchés.

Un exemple de structuration obtenue par ce moyen est celui du taxème { *pouvoir*, *autorité*, *gouvernement* } dans le cadre du thème « territoire » : les auteurs font apparaître que *gouvernement* se distingue des deux autres mots par son aspect institutionnel et structuré, *autorité* supposant en outre par rapport à *gouvernement* un champ d'action restreint à un domaine donné ; *pouvoir* comporte pour sa part un élément d'instabilité, et est défini comme plus « changeant » qu'*autorité*. Ces résultats paraissent très prometteurs pour la construction de lexiques, et permettent aux auteurs d'effectuer une structuration pertinente d'un taxème par des sèmes spécifiques. L'automatisation reste néanmoins limitée à la construction de listes de mots-indices dont il revient à un expert de décider lesquels sont pertinents, et quelles nuances ils définissent. Ce rôle très important dévolu à l'interprétation manuelle présente le défaut majeur de ne pas garantir l'endogénéité de l'analyse, et de permettre à l'interprétant d'introduire dans celle-ci nombre de présupposés non inhérents aux textes étudiés. Ainsi est-il avancé dans [PS99] que les mots { *ministre*, *union*, *formation* } sont révélateurs de l'aspect institutionnel et structuré de *gouvernement*, tandis que { *place*, *prise*, *arrivée*, *année* } suggèrent l'instabilité de *pouvoir*, sans que puissent être rendues manifestes les raisons justifiant la différence d'évaluation ainsi suggérée entre *formation* (« d'un gouvernement ») et *arrivée* (« au pouvoir »).

Afin de tenter de contourner le biais interprétatif introduit par la méthodologie proposée dans [PS99], nous avons donc étudié au cours de nos recherches les possibilités d'automatisation de celle-ci, commençant pour cela par constater de nous-même, dans un premier temps, la teneur des indices rassemblés par une étude de voisinage immédiat, et la difficulté qu'il peut y avoir à en extraire manuellement des informations de distinction de sens. Ne travaillant pas sur des données identiques (quoique issues d'un corpus similaire), nous avons fait le choix d'étudier un exemple radicalement différent, et de nous limiter pour simplifier l'analyse à un simple couple de mots. Nous présentons donc à la section suivante les résultats d'une étude manuelle visant à distinguer le sens de *client* de celui de *consommateur*.

4.4.2. Étude de cas : distinction client / consommateur

Par souci de clarté et de simplification, nous nous limitons pour la comparaison présentée ici entre *client* et *consommateur* à l'étude des adjectifs employés dans le sous-corpus « nouvelles technologies » (qui nous sert toujours de référence) comme épithètes à ces deux mots ; les recherches que nous avons menées par ailleurs montrent que les résutats obtenus en prenant en compte des indices de catégories morphosyntaxiques différentes suggèrent des conclusions similaires. Les indices contextuels étudiés sont extraits selon un critère très simple consistant à retenir les adjectifs soit immédiatement contigus à une occurrence du mot à caractériser, soit apparaissant après celle-ci et séparés d'elle par un adverbe. Le tableau de la figure 4.4 donne pour les deux mots considérés la liste des adjectifs répondant à ce critère de sélection.

Les adjectifs rassemblés ont été regroupés manuellement en 11 classes sémantiques dont le tableau précise la dénomination. En effet, quand les auteurs de [PS99] calculent afin de mettre au jour les différences d'emploi des mots la différence entre leurs ensembles caractéristiques, puis effectuent des regroupements au sein de l'ensemble résultant de cette opération, la procédure que nous adoptons consiste à regrouper tout d'abord les indices contextuels en classes, puis à rechercher les différences entre celles-ci : cette approche, sensiblement équivalente dans ses résultats, présente l'avantage d'être moins sensible à la variation lexicale ; elle nous semble en outre permettre un exposé plus clair.

Des catégories sémantiques d'adjectifs exposées dans le tableau de la figure 4.4, seule la dernière est propre à *client* à l'exclusion de *consommateur*. L'interprétation de cette différence est délicate : on ne peut affirmer *a priori* que *client* soit valorisé ou non par rapport à *consommateur*, mais seulement qu'il est plus « sujet à jugement ». Un axe de différentiation est ainsi mis au jour, que nous sommes toutefois encore réticent à qualifier de sème spécifique tant cette information semble « sous-spécifiée » en l'absence d'expression d'un « sens » le long de cet axe. Il est également possible de découvrir des différences au

	client	consommateur
1. adjectifs « contingents » sans lien intuitif particulier avec le sens du mot	**final**, autre, fort, significatif, politicien, mineur, actuel, nouveau, respectif, bénéficiaire	**final**, quelconque, concret, authentique, remboursé, atomisé, souverain
2. dénombrement	**seul**, **nombreux**, **premier**, multiple, deuxième, cinquième, quatrième, second, sixième, quinzième, troisième	**seul**, **nombreux**, **premier**
3. importance	**gros**, **important**, **principal**, **grand**, significatif	**gros**, **important**, **principal**, **grand**
4. nationalité	**européen**, **français**, **soviétique**, **américain**, **occidental**, **asiatique**, russe, chinois, latino-américain, belge, afghan, nord-américain, catalan, maronite, étranger,	**européen**, **français**, **soviétique**, **américain**, **occidental**, **asiatique**, mondial, planétaire, israélien, japonais, algérien, hongrois, néerlandais, africain, francophone
5. condition économique	**aisé**, **solvable**, **pauvre**, riche, privilégié, démuni, humble, insolvable, fortuné, indigent	**aisé**, **solvable**, **pauvre**, impécunieux, modeste
6. qualification « objective »	**petit**, **local**, simple, civil, collectif, public, privé	**petit**, **local**, urbain, individuel, domestique, salarié, solitaire
7. qualification plus « subjective »	sérieux, honnête, respectable, illustre, réputé, vulgaire, traditionnel	fameux, qualifié, stimulé, dynamique, ignorant
8. adjectifs reflétant le statut de « proie » recherchée	**éventuel**, **potentiel**, désirable, indésirable, pressenti, indispensable	**éventuel**, **potentiel**,
9. habitude, aspect répétitif du statut de client ou consommateur	**régulier**, **occasionnel**, **nouveau**, **ancien**, quotidien, fidèle, captif, habituel, fixe, docile	**régulier**, **occasionnel**, **nouveau**, **ancien**
10. attitude émotionnelle	désireux, exigeant, impatient, imprévisible, médusé, extasié, réticent, récalcitrant, mécontent, indécis, satisfait, favorable	avide, boulimique, inquiet, culpabilisé, disposé
11. évaluation / jugement	bon, mauvais, médiocre, excellent, piètre, cher, précieux	

FIG. 4.4 – *Adjectifs voisins des noms client et consommateur dans le sous-corpus « nouvelles technologies », regroupés manuellement en catégories sémantiques. Les adjectifs marqués en gras sont ceux partagés par les deux mots.*

sein des autres classes d'adjectifs : si nous comparons dans les attitudes émotionnelles associées à *client* et *consommateur* les adjectifs révélant une forme de désir ({*désireux*, *exigeant*} pour le premier, {*avide*, *boulimique*} pour le second), *client* paraît plus sensé, moins « animal » que *consommateur*. Il bénéficie en outre d'un ensemble de « qualifications subjectives » plus étendu, et plus valorisant.

Nous pouvons déduire plus directement une distinction établie entre *client* et *consommateur* dans notre sous-corpus d'étude en observant les phrases où ils apparaissent de manière conjointe. Sur l'ensemble du corpus, trois phrases répondent à ce critère :

– « Cette efficacité pourrait bien être surpassée par la formidable extension, promise par Internet, du marketing direct, c'est-à-dire du ciblage des **consommateurs** à travers des relevés de leurs achats et de leurs préférences médiatiques ; les fabricants de produit de consommation élaborent consciemment des sites leur permettant d'établir de nouveaux rapports avec le **consommateur**, d'engager une conversation continue avec chaque **client** désirable. »
– « Plus d'un milliard d'êtres humains tu plaisantes : plus d'un milliard de **consommateurs** potentiels, de **clients** à satisfaire. »
– « À l'autre extrémité du marché, les entreprises qui s'adressent au **consommateur** impécunieux doivent également bien s'en sortir ; Painwebber attire ainsi l'attention de son **client** sur les sociétés Wall-Mart et Costco, spécialisées dans les très grandes surfaces bas de gamme, un genre qui, grâce aux tarifs qu'il pratique, continue à gagner des parts de marché. »

La distinction est cette fois plus clairement remarquable : *client* est nettement valorisé par rapport à *consommateur*. En effet, le premier, « client désirable », « à satisfaire », est acteur d'une « conversation » avec le vendeur ; le consommateur, objet d'un « ciblage », n'appartient d'ailleurs pas au même monde : les clients de l'agence de conseils en investissements boursiers Painwebber n'ont que peu à voir avec les « consommateurs impécunieux » fréquentant Wall-Mart. Selon cette analyse, *client* et *consommateur* se trouvent donc séparés sur une échelle évaluative, le premier bénéficiant d'un trait mélioratif[4].

L'analyse présentée dans cette sous-section illustre la relative facilité avec laquelle un interprétant humain peut découvrir parmi les caractérisations contextuelles des usages de deux mots des indices exprimant les distinctions de sens existant entre eux — en particulier si sa compréhension des mots comparés et sa connaissance des « codes » en vigueur dans les textes considérés lui

[4]Il est à noter que cette distinction relève plus d'une norme propre au corpus étudié que du système de la langue à proprement parler, en d'autres termes qu'elle est *afférente* plutôt qu'*inhérente* (la distinction « officielle » entre client et consommateur étant que le premier est en relation avec un fournisseur, et le second avec un produit). Le parti pris de guidage exclusif par le texte nous conduit néanmoins en l'état de nos recherches à ignorer la distinction inhérent / afférent.

suggère *a priori* une intuition de ce que peuvent être ces distinctions... Nous étudions maintenant quelles sont les possibilités d'automatisation de ce processus interprétatif.

4.4.3. Tentatives d'automatisation

Au cours de l'analyse manuelle du cas *client* / *consommateur*, nous avons fait usage avec une efficacité plus ou moins grande de trois méthodes :

- regroupement des indices en classes sémantiques afin de rechercher si l'une ou plusieurs d'entre elles sont uniquement présentes dans les voisinages d'un des deux mots,
- étude des distinctions de composition entre classes sémantiques d'indices « correspondantes »,
- réduction du domaine d'étude aux phrases rassemblant les deux noms comparés.

Nous décrivons rapidement dans cette section les tentatives d'automatisation de ces trois outils que nous avons réalisées — ou, plus exactement, les obstacles s'opposant à leur automatisation. Nous commençons cette étude par le dernier des trois outils cités, qui relève de la sélection des données à employer plus que de leur analyse.

4.4.3.1. Sélection des énoncés pertinents pour la distinction

L'intuition ayant motivé notre choix de restreindre l'analyse du texte au phrases où les deux mots étudiés apparaissent conjointement est que, ces mots étant de sens très proches, ils sont plus susceptibles, dans le cas où ils apparaissent à proximité l'un de l'autre, de voir leur sens précisé, leur spécificité explicitée. Dans le cas de la paire *client* / *consommateur*, cette idée est largement confirmée, et une simple mise en opposition des quelques adjectifs qualifiant ces deux mots dans les énoncés retenus donne de la distinction qui les oppose une représentation très aisément interprétable.

Malheureusement, ce principe ne s'applique en général que très mal : deux mots apparaissant au sein d'un même taxème peuvent, suivant les cas, ne jamais apparaître dans la même phrase ou, au contraire, avoir des centaines de telles cooccurrences. De plus, les distances sémantiques entre mots au sein des taxèmes construits au chapitre précédent sont très variables : en effet, en construisant des sous-corpus d'étude monothématiques, nous rassemblons des textes particulièrement adaptés pour définir le sens des mots « centraux » du domaine correspondant, permettant ainsi la constitution de taxèmes aussi cohérents que par exemple {*ordinateur*, *micro-ordinateur*} dans le domaine des « nouvelles technologies ». Les mots plus « annexes », en revanche, tendent à être moins bien qualifiés par les textes considérés, et peuvent former des taxèmes de sens beaucoup plus général, comme (toujours dans le même do-

maine) {*laser*, *navette*, *engin*}. Dans ce second cas, l'apparition conjointe de deux mots dans une même phrase ne pose à son auteur aucun problème de *distinguo*, ce qui met à mal l'hypothèse de travail que nous avons posée en introduction de cette sous-section.

Nous sommes donc contraint pour toute analyse à ne pas nous limiter à prendre en compte d'hypothétiques phrases contenant conjointement les mots à comparer, mais à considérer la totalité des informations de voisinages les concernant. Afin de structurer cette somme d'informations pour la rendre exploitable, nous étudions maintenant les possibilités de classification automatique des mots de ces voisinages.

4.4.3.2. Regroupement en classes des indices

Le regroupement en classe des adjectifs propres aux voisinages respectifs des noms comparés est indispensable pour leur donner une intelligibilité globale. Il nous permet au cours de notre étude manuelle de repérer ceux dont le sens général est commun aux deux noms, et de mettre en évidence par défaut d'une telle correspondance une dimension selon laquelle *client* se démarque de *consommateur*, celle de l'évaluation — même si nous ne pouvons dans un premier temps affirmer dans quelle « direction » suivant cet axe se fait la distinction.

Nous avons tenté d'adapter à la classification des adjectifs ainsi regroupés les méthodes de constitution de classes sémantiques développées au chapitre précédent, soit en analysant séparément les ensembles d'adjectifs extraits pour les deux mots comparés, soit en les considérant conjointement afin de déduire de la classification les correspondances éventuelles entre classes construites dans les deux ensembles, soit en ne considérant que leurs différences ou intersection, suivant la méthodologie de [PS99]. Que nous employions pour ce faire la technique développée pour prendre en compte l'intégralité des données du corpus (présentée à la section 3.4) ou celle ne travaillant que sur un extrait thématique de celui-ci[5] (présentée à la section 3.5), les classifications produites ne permettent d'organiser qu'une part relativement faible des adjectifs regroupés, ceci étant en particulier dû au très faible nombre d'occurrences d'une grande majorité d'entre eux. En particulier, aucune des classifications des adjectifs caractérisant *client* que nous avons pu construire ne fait clairement apparaître de classe d'adjectifs évaluatifs telle que celle que nous avons générée manuellement. Si même il était possible de faire apparaître cette classe parmi les adjectifs associés à *client*, resterait à satisfaire la nécessité encore plus problématique d'affirmer fermement qu'il n'est pas possible de lui trouver une contrepartie parmi les adjectifs caractérisant *consommateur*.

Nous trouvant dans l'impossibilité de construire de manière fiable des

[5]L'association *client* / *consommateur* étant relativement « universelle », nous avons pu dans ce cas expérimenter sur divers thèmes.

classes sémantiques parmi les indices qualifiant les deux mots, la question de savoir s'il est possible de mettre en évidence de manière automatique des différences sensées entre leurs compositions reste purement théorique, et d'un intérêt limité. Une telle entreprise supposerait en outre d'être à même d'identifier des nuances sémantiques entre adjectifs, autrement dit d'avoir une connaissance des sèmes spécifiques les distinguant, problème que nous n'avons pas abordé, nous limitant à l'étude des noms. C'est pourquoi nous ne nous attardons pas plus sur cette direction de recherche.

La faiblesse des résultats observés lors de ces premières expériences s'explique en partie par un problème d'insuffisance de données : nous avons pu constater au chapitre précédent que les méthodes de regroupement de mots en classes sémantiques que nous avons développées sont, comme toute méthode d'analyse statistique de données, assez sensibles à la fréquence des mots étudiés. Les adjectifs apparaissant dans les voisinages des mots classés pouvant présenter des nombres d'occurrences très variables, il n'est pas surprenant qu'ils ne se prêtent que mal aux techniques de classifications statistiques. Les raisons des difficultés que nous rencontrons dans notre recherche sont néanmoins également théoriques, comme nous le montrons maintenant.

4.4.4. Une voie sans issue ?

Au vu des résultats mitigés de l'exploitation d'indices de voisinages immédiats pour la recherche de sèmes spécifiques, on est en droit de s'interroger sur la pertinence de ces données pour la tâche entreprise. Cette section présente les enseignements qu'une étude proprement linguistique de la question peut apporter à ce sujet. Nous y présentons dans un premier temps une typologie des isotopies pouvant se développer dans cet énoncé minimal qu'est la paire nom-adjectif, puis exposons les résultats d'une étude visant à évaluer les importances respectives (en volume) de ces différents types d'isotopies dans le corpus sur lequel nous travaillons.

4.4.4.1. Typologie des isotopies dans les groupes nom-adjectif

Étudiant « pour l'intérêt du cas limite » les conditions minimales dans lesquelles une isotopie peut se développer, F. Rastier est amené au chapitre VI de [Ras96] à considérer en détail la relation nom-adjectif. Il propose une typologie des isotopies pouvant exister dans un tel contexte élémentaire, dont nous ne présentons ici qu'une version simplifiée adaptée à nos besoins.

1. Isotopie générique simple Dans ce cas, nom et adjectif ont en commun un sème générique de plus ou moins grande généralité. Il s'agit au minimum d'un sème macrogénérique (par exemple / animé / dans « une chatte joueuse »), condition en-deçà de laquelle on se trouve en situation d'allotopie* (absence

d'isotopie) : « une montagne joueuse ». Il peut également s'agir d'un sème mésogénérique (/ mécanique / dans « une voiture puissante ») ou microgénérique (/ véhicule léger / dans « une voiture décapotable »).

2. Isotopie générique et spécifique Dans ce cas, nom et adjectif sont en situation d'isotopie générique et, de plus, l'adjectif ne fait qu'expliciter une spécificité du nom. C'est ce qu'on appelle couramment un pléonasme, comme « panacée universelle » ou « dilemme indécidable ».

3. Isotopie spécifique-générique Un sème spécifique du nom est également présent, mais en tant que sème générique, dans l'adjectif. C'est par exemple le cas dans « abattement fiscal », où / fiscal / est précisément le sème spécifique qui distingue *abattement* de *réduction*, *rabais*, *ristourne*, *etc.* C'est ce cas de figure particulier qui nous intéresse pour le type de recherche de sèmes spécifiques dans le voisinage immédiat présenté à la section 4.4.2.

4. Isotopie générique-spécifique À l'inverse du cas précédent, un sème générique du nom est présent en tant que sème spécifique dans l'adjectif. Il s'agit de l'interprétation donnée par F. Rastier du principe de solidarité lexicale introduit par E. Coseriu [Cos68], dont un exemple classique est celui des adjectifs anglais *pretty* et *handsome* qui, appliqués à des êtres humains, se voient porteurs (relativement à un taxème que nous nommerons simplement // joli //) de sèmes spécifiques féminin pour le premier, masculin pour le second (indépendamment de toute notion de genre grammatical, inexistant pour les adjectifs en anglais). Pour reprendre les termes de F. Rastier à ce sujet ([Ras96], p.76), « *handsome girl* ne peut désigner qu'une fille hommasse, voire un personnage plus équivoque encore ».

5. Hypallage Nom et adjectif sont apparemment en allotopie, mais l'isotopie est assurée *via* un troisième mot, présent ou non en contexte : F. Rastier cite l'exemple littéraire « boulot transpirant » (c'est le travailleur, sous-entendu, qui transpire), nous pouvons pour notre part relever de manière plus pragmatique dans *le Monde diplomatique* « une analyse concrète de la situation », où seules la situation et les données supposées de l'analyse peuvent raisonnablement être concrètes.

6. Allotopie Absence totale d'isotopie ou incompatibilité de sèmes. F. Rastier en distingue plusieurs types (de « un cercle carré » à « une vache gazeuse ») sur lesquels nous ne nous attardons pas, cette catégorie n'étant que peu pertinente dans les textes que nous étudions.

Type	Compte	Proportion
1. Générique simple	512	0,84
dont : macrogénérique	237	0,38
mésogénérique	190	0,32
microgénérique	85	0,14
2. Pléonasme	3	0,01
3. Spécifique-générique	45	0,07
4. Générique-spécifique	42	0,07
5. Hypallage	7	0,01
6. Allotopie	0	0,0
Total	608	1,0

TAB. 4.1 – *Répartition des types d'isotopies présents entre les paires nom-épithète extraites d'un extrait aléatoire de 100 paragraphes du corpus du* Monde Diplomatique

De ces six catégories, seule la troisième peut occasionner l'expression dans le texte d'un sème spécifique du nom sous forme générique dans son contexte immédiat, et présente donc un intérêt pour notre recherche. L'exploitation de cette relation pour la mise au jour automatique de sèmes spécifiques est néanmoins soumise à la condition qu'elle soit suffisamment représentée dans les textes pour permettre une analyse numérique. Nous présentons dans ce qui suit les résultats d'une expérience que nous avons réalisée afin d'évaluer ce facteur de volume.

4.4.4.2. Recensement des types d'isotopies nom-adjectif

Dans le but d'étudier les importances relatives des différents types d'isotopies décrits, nous avons extrait tous les liens entre noms et adjectifs épithètes apparaissant dans une sélection aléatoire de 100 paragraphes du corpus du *Monde diplomatique*, puis attribué manuellement aux 608 paires ainsi extraites l'une des catégories proposées par F. Rastier. Nous avons au cours de ce travail respecté la méthodologie de la sémantique interprétative, faisant systématiquement appel, afin de préciser le référentiel de définition des sémèmes et sèmes, au contexte d'apparition des paires étudiées. Ne prétendant néanmoins pas à une quelconque expertise linguistique, nous ne présentons les résultats de ce décompte qu'à titre d'ordre de grandeur.

Les résultats rassemblés dans le tableau 4.1 suggèrent que la très grande majorité des isotopies en effet présentes dans le texte relèvent de la première catégorie, seuls 7 % des paires recensées se trouvant dans un rapport séman-

tique exploitable pour la recherche de verbalisation de sèmes spécifiques. Cette faible densité d'information pertinente laisse peu d'espoir pour l'automatisation de la recherche de sèmes spécifiques par l'étude des liens noms-adjectifs : il est en effet impossible dans les conditions constatées de se passer d'une étape d'analyse humaine pour la sélection et l'organisation des indices réellement intéressants.

La situation est sensiblement identique dans le cas des formulations de type « N de N »[6] : d'après une étude réalisée selon les mêmes critères (mais en n'employant cette fois qu'une sélection de 30 paragraphes de notre corpus), plus de 80 % des paires de mots étudiées se trouvent dans une relation relevant de la première catégorie (générique-générique), seuls 9 % des cas examinés correspondant à une relation de type spécifique-générique.

Bien que nous n'ayions pas eu le loisir d'analyser rigoureusement le cas des relations noms-verbes, et que notre « sondage » des relations nom-nom reste superficiel, cette rapide étude suggère que les sèmes spécifiques semblent beaucoup moins impliqués que les sèmes génériques dans les relations syntagmatiques immédiates, ce qui rend problématique leur extraction à partir de voisinages proches. C'est pourquoi nous nous tournons à la section suivante vers l'exploitation de phénomènes d'isotopie « à longue portée », qui permettent en effet d'obtenir des résultats nettement plus satifaisants.

4.5. Exploitation d'isotopies à « longue distance »

Le principe de la méthode présentée dans cette section, qui permet une automatisation partielle de l'acquisition sur corpus d'informations de distinctions entre sens de mots au sein d'un taxème, est de considérer pour cette tâche non plus les indices apparaissant dans le texte à proximité immédiate des mots à comparer, mais au contraire ceux qui, tout en participant à leur entour linguistique, se trouvent dans une position plus éloignée. Ce choix, que nous avons introduit à la section 4.4.4 par une étude « de volume », se justifie également d'un point de vue plus théorique. En effet, ce qui est en jeu au niveau textuel du voisinage proche est la cohésion des énoncés, l'existence de relations assurant le « flux » du texte ; cette exigence d'identité, d'équivalence, implique naturellement un rôle prédominant des sèmes génériques. Les sèmes spécifiques, pour leur part, traduisent à l'échelle microsémantique la nuance, l'opposition, la gradation, *etc.* — autant de notions relevant de la construction d'idées, de la dialectique, de l'argumentation. Il semble donc pertinent de s'intéresser pour la recherche de sèmes spécifiques aux isotopies développées à l'échelle du discours, plutôt qu'aux isotopies « minimales » dont faisaient usage les expériences déjà présentées.

[6] Plus précisément, nous nous sommes intéressé aux formulations définies par le patron général N1 de <déterminant>? <adjectif ou adverbe>* N2.

La technique mise au point afin d'analyser ces données pour en faire émerger une organisation interprétable revient au principe de caractérisation implicite des sèmes spécifiques, en rassemblant les paires de mots dont les usages se distinguent par des éléments de contexte similaires (section 4.5.1). Les propositions mises en avant par la méthode développée requérant toujours une part de filtrage et d'interprétation manuelle, nous présentons à la section 4.5.2 la procédure que nous avons mise au point pour réaliser cette opération en minimisant les possibilités d'introduction par l'« expert » de connaissances non inhérentes au texte, avant de présenter les résultats définitifs (section 4.5.3).

4.5.1. Rapprochement de paires de mots distingués par des spécialisations similaires

Conformément au principe adopté de travail « à longue portée », nous considérons donc pour chaque occurrence des noms à caractériser un voisinage constitué par la totalité des mots de la phrase où elle apparaît[7], à l'exception de ses voisins les plus proches (dans une fenêtre d'exclusion de trois mots à droite et à gauche de chaque occurrence). Les catégories de mots retenues pour prendre part à la caractérisation sont les noms, verbes et adjectifs.

Il résulte de ce choix, considérant un entour linguistique aussi étendu que la phrase, que les indices retenus pour chacun des noms considérés constituent des ensembles de tailles très importantes, probablement très bruités par le fait qu'un mot donné ne participe pas simultanément à toutes les isotopies développées dans la phrase où il apparaît. Il semble donc que peu d'espoir puisse être placé dans la recherche directe de désignations explicites de sèmes spécifiques dans ces ensembles d'indices. C'est pourquoi nous nous tournons pour cette étude vers la construction d'une représentation implicite de sèmes spécifiques telle que présentée à la section 4.3.1, par rassemblement des paires de mots dont les éléments respectifs sont distingués par des nuances de sens similaires. Afin de maximiser les chances de voir s'opérer de tels rapprochements, il est naturellement nécessaire de ne plus s'intéresser à la structuration d'un taxème unique comme nous le faisions précédemment : plus les paires de mots étudiées sont nombreuses, plus les chances de découvrir parmi elles des ensembles représentatifs de sèmes spécifiques augmentent. C'est pourquoi nous considérons simultanément tous les taxèmes construits pour un domaine donné grâce aux traitements présentés au chapitre 3, et cherchons des points communs entre toutes les paires de mots telles que leurs deux éléments appartiennent à un même taxème. Nous noterons dans la suite $\mathcal{T}$ l'ensemble des taxèmes $T_1, \ldots, T_n$. M désigne l'ensemble des mots rassemblés par les T_i.

[7] F. Rastier émet de nombreuses réserves concernant la pertinence de la phrase comme échelle d'étude pour l'analyse textuelle, mais nous sommes contraint d'y faire appel à défaut de pouvoir définir simplement une limite plus adaptée.

À chaque mot m de M, nous associons l'ensemble I_m de tous les noms, verbes et adjectifs apparaissant dans les mêmes phrases que m mais éloignés de lui par une distance minimale de 3 mots (valeur indépendante des longueurs des phrases où apparaît le mot, définie empiriquement afin de refléter notre volonté d'ignorer le voisinage immédiat du mot). Si plusieurs mots d'une même classe sont simultanément présents dans une même phrase, nous n'associons pas tous les mots de la phrase à chacun des mots, puisque notre but est de faire apparaître autant que possible des différences d'usage entre eux. Chaque mot-indice est donc rattaché à la caractérisation du mot qui est le plus proche de lui, en nombre de positions, dans la phrase[8].

Les objets de cette étude sont les paires de mots (m_1, m_2) telles que m_1 et m_2 appartiennent à un même taxème T_i. Notre objectif est de définir une mesure de similarité dont la valeur soit d'autant plus élevée que, pour deux paires (m_1, m_2) et (m'_1, m'_2), les différences d'usages distinguant m_1 de m_2 sont similaires à celles distinguant m'_1 de m'_2. Afin de caractériser cette différence d'usage, nous associons à chaque paire (m_1, m_2) une paire d'ensembles (E_1, E_2) définis par $E_1 = I_{m_1} \setminus I_{m_2}$ et $E_2 = I_{m_2} \setminus I_{m_1}$. E_1 contient donc l'ensemble des indices contextuels exclusivement associés à m_1, et E_2 ceux exclusivement associés à m_2.

La similarité entre deux paires de mots est calculée en plusieurs étapes, et fait usage de la « distance sémantique » entre mots calculée au chapitre précédent (section 3.4) en employant la totalité des données textuelles du corpus. Si nous notons d cette distance et travaillons sur deux paires de mots (m_1, m_2) et (m'_1, m'_2), respectivement caractérisées par les paires d'ensembles (E_1, E_2) et (E'_1, E'_2), nous définissons tout d'abord un « cardinal d'intersection floue » cif par :

$$cif\,(E_1, E'_1) = \frac{1}{2}\left[\sum_{m_1 \in E_1} \max_{m_2 \in E'_1} (1 - d\,(m_1, m_2)) + \sum_{m_2 \in E'_1} \max_{m_1 \in E_1} (1 - d\,(m_2, m_1))\right] \tag{4.1}$$

Les valeurs de d étant normalisées pour appartenir à l'intervalle $[0, 1]$, cette formule peut être comprise comme le cardinal de l'intersection de deux « ensembles flous » F_1 et F'_1 tels que le degré d'appartenance d'un mot à F_1 (resp. F'_1) soit égale à sa proximité $(1-d)$ maximale avec l'un des mots de E_1 (resp. E'_1). Nous l'employons pour calculer la similarité ς entre deux ensembles de mots-indices E_1 et E'_1 :

[8] Pour être complet dans notre description de la procédure mise en œuvre, précisons qu'en cas d'égalité de distance, le mot-indice est affecté par souci égalitaire au moins fréquent des deux mots « en concurrence ». Ce choix a un impact tout à fait marginal sur les résultats observés.

$$\varsigma\left(E_1, E_1'\right) = \frac{cif\left(E_1, E_1'\right)}{\mathrm{Card}\left(E_1\right) + \mathrm{Card}\left(E_1'\right) - cif\left(E_1, E_1'\right)} \tag{4.2}$$

ς n'est autre qu'un indice de Jaccard adapté aux cas d'ensembles flous afin de prendre en compte l'information fournie par d. Nous calculons enfin la similarité s entre les deux paires de mots considérées en réalisant la moyenne des similarités observées entre leurs ensembles d'indices E_1 et E_1' d'une part, E_2 et E_2' d'autre part :

$$s\left(\left(m_1, m_2\right), \left(m_1', m_2'\right)\right) = \frac{\varsigma\left(E_1, E_1'\right) + \varsigma\left(E_2, E_2'\right)}{2} \tag{4.3}$$

Les mesures de similarité entre paires de mots ainsi calculées sont normalisées par la méthode de centrage et réduction ligne par ligne et colonne par colonne présentée à la section 3.4.3. Une exploration manuelle des rapprochements entre paires suggérés par cette mesure montre l'existence de regroupements « intéressants », faisant sens par rapport à notre objectif. Néanmoins, toutes nos tentatives d'exploitation directe de ces valeurs de similarité par des méthodes de classification classiques — par exemple la construction d'un arbre de classification dont les feuilles seraient les paires de mots étudiées — mènent à une perte d'information importante par rapport aux rapprochement pertinents qu'une analyse manuelle des similarités permet de mettre au jour. Nous avons donc défini une procédure simple pour faciliter l'exploitation manuelle de ces résultats en évitant autant que possible l'« enrichissement » de ceux-ci par des connaissances non intrinsèques aux textes considérés.

4.5.2. Procédure d'exploitation manuelle des résultats

À chacune des paires de mots considérées est associée la liste des 5 paires présentant avec elle une similarité maximale. Le processus de filtrage et interprétation manuelle des résultats est réalisé en deux étapes à partir de ces listes de paires les plus similaires que nous nommerons dans la suite plus simplement « listes de connexions » de chaque paire.

Sélection des liens « évidents » En parcourant rapidement les listes de connexions ainsi produites, nous avons relevé les associations entre paires les plus immédiatement marquantes selon un critère d'analogie (« a_1 est à b_1 ce que a_2 est à b_2 »), et associé à chacune de ces associations une dénomination reflétant la relation apparente entre les extrémités des deux paires ainsi rassemblées. Travaillant sur l'ensemble de taxèmes du domaine « nouvelles technologies » dont le contenu est précisé à la figure 4.5, les associations de paires de mots ainsi relevées sont celles présentées dans le tableau 4.2. Sur chaque ligne du tableau sont reportées les deux paires rassemblées ainsi que la dénomination

Paires regroupées	Nuance
(imprimerie / photographie) / (télécopieur / télévision)	/ texte - image /
(cerveau / intelligence) / (téléphonie / téléphone)	/ condition /
(autoroute / réseau) / (transport / communication)	/ concret - abstrait /
(mastodonte / géant) / (rachat / fusion)	/ agressif - consensuel /
(téléphonie / télécommunication) / (radio / télévision)	/ ancien - moderne /
(technique / technologie) / (câble / satellite)	/ simple - sophistiqué /
(présentation / lecture) / (éditeur / lecteur)	/ production - consommation /

TAB. 4.2 – *Paires de mots rapprochées sur le domaine « nouvelles technologies » par une étude manuelle superficielle des « listes de connexions » calculées automatiquement*

extrapolée de la nuance distinguant leurs membres (que nous notons en adoptant la convention de représentation des sèmes, entre barres obliques) : sur la première ligne, par exemple, *imprimerie* se distingue de *photographie*, tout comme *télécopieur* de *télévision*, par le fait que les premiers termes de ces paires sont liés au monde du texte, et les seconds à celui de l'image. Sur la deuxième ligne, *cerveau* est une condition d'existence de *intelligence*, tout comme l'existence de la technologie *téléphonie* est nécessaire à celle du *téléphone*. Certaines des associations présentées peuvent paraître moins intuitives que d'autres ; nous discutons de ces choix de manière plus approfondie à la section suivante, lors de la présentation des résultats complets.

Validation des rapprochements La validation à proprement parler des résultats est assistée par un système automatisé au fonctionnement assez simple : pour chaque paire de mot étudiée p, il construit l'ensemble des « noms de nuances » définis à l'étape précédente qui ont déjà été validés pour une au moins des paires présentes dans la liste de connexions de p, puis propose au « valideur » de confirmer ou infirmer la pertinence de ces nuances pour caractériser la différence de sens distinguant les éléments de p. Par exemple, si la paire (*imprimerie* / *photographie*) fait partie de la liste de connexions de p, le valideur devra évaluer la validité de la nuance / texte - image / pour distinguer les deux mots de p. Les associations validées sont prises en compte par le système afin de générer d'éventuelles nouvelles propositions impliquées par ces nouvelles connaissances. Le travail de validation se termine lorsque toutes les possibilités ont été épuisées.

L'ensemble des deux étapes décrites correspond à une heure de travail humain environ pour étudier les propositions faites sur un ensemble de classes sémantiques telles que celles de la figure 4.5. Nous présentons maintenant les résultats complets auxquels elles permettent d'aboutir.

{ surveillance, observation, investigation, enregistrement }
{ journal, feuilleton, émission, spot }
{ système, programme, dispositif }
{ équipement, matériel, infrastructure }
{ journaliste, éditeur, lecteur, rédacteur, présentateur }
{ informatique, électronique, biologie }
{ usage, utilisation, maîtrise }
{ major, industriel, multinationale }
{ groupe, atelier, corporation, câblo-opérateur, entreprise, opérateur, firme, compagnie }
{ client, consommateur, utilisateur }
{ presse, quotidien, média, magazine }
{ transparence, disponibilité, fiabilité }
{ mondialisation, déréglementation, globalisation, dérégulation }
{ communication, navigation, transport, transmission }
{ télécommunication, téléphonie, téléphone }

{ recherche, expérimentation }
{ information, donnée }
{ désinformation, relations publiques }
{ performance, rentabilité }
{ télévision, radio }
{ satellite, câble }
{ adresse, hypertexte }
{ cédérom, cd-rom }
{ planète, globe }
{ vie privée, confidentialité }
{ fichier, index }
{ photographie, imprimerie }
{ usager, pirate }
{ signal, paquet }
{ distance, rayon }
{ lecture, présentation }
{ ordinateur, micro-ordinateur }
{ engin, navette, laser }
{ mutation, innovation }
{ tarif, frais }
{ réseau, autoroute }
{ courrier, messagerie }
{ flux, circulation }
{ chaîne, station }

{ concentration, densité }
{ audience, tirage }
{ noeud, maillage }
{ clavier, interface }
{ intelligence, cerveau }
{ pionnier, pilote }
{ univers, paysage }
{ fabricant, constructeur }
{ technologie, technique }
{ vitesse, précision }
{ échelle, dimension }
{ norme, standard }
{ distribution, commercialisation }
{ diffusion, réception }
{ manipulation, conditionnement }
{ connexion, interconnexion }
{ loisir, attraction }
{ téléviseur, télécopieur }
{ logiciel, machine }
{ concurrence, concurrent }
{ géant, mastodonte }
{ fusion, rachat }
{ cassette, copie }

FIG. 4.5 – *Taxèmes issus des traitements présentés au chapitre 3 employés pour notre expérience de recherche de sèmes spécifiques (la séparation apparente en deux parties n'est due qu'à un impératif de mise en page)*

4.5.3. Résultats

Cette section détaille, pour chacun des sèmes spécifiques mis au jour et manuellement nommés selon la procédure décrite ci-dessus, l'ensemble des paires de mots qu'il permet de distinguer. Nous y discutons également du sens qu'il est possible de donner à ces sèmes au-delà de leur simple nommage, nécessairement approximatif, et précisons les raisons qui nous ont poussé à retenir certaines associations parfois peu évidentes.

concret - abstrait La première nuance retenant notre attention est celle que nous avons désignée par l'alternance / concret - abstrait /. C'est elle qui distingue selon notre analyse le plus grand nombre de paires de mots, comme le montre la figure 4.6. Ce volume important s'explique en particulier par le fait que se trouvent mêlés ici deux types de distinctions : certaines paires de mots font référence à des concepts dont l'un est plus concret, matériel que l'autre (par exemple *cerveau* / *intelligence*, *dispositif* / *système* ou *dimension* / *échelle*[9]), d'autres à des situations où c'est le terme lui-même qui est plus concret (c'est-à-dire plus spécifique, mais aussi plus « terre-à-terre »). Ainsi, dans le cas de la distinction entre *opérateur* et *entreprise*, le premier est à la fois plus précis et plus concret (un *opérateur* est une entreprise ayant une activité précise, identifiable et observable) que le second.

Ce qui apparaît ici est un point de convergence entre les sèmes spécifiques que nous recherchons et la notion lexicale traditionnelle d'hyperonymie, sans qu'il soit possible ni par les indices que nous extrayons des données textuelles, ni même parfois par l'interprétation humaine, d'établir une distinction nette entre la distinction / concret - abstrait / que nous mettons en avant et l'alternative « spécifique / général » (entre hyponyme et hyperonyme) qui vient s'y agréger. La ligne directrice que nous avons adoptée au cours de la validation manuelle consiste à ne retenir la distinction / concret - abstrait / pour les paires en relation d'hyponyme à hyperonyme que si le premier mot peut en effet s'interpréter de manière plus « matérielle » que le second (comme c'est le cas pour la paire *opérateur* / *entreprise*).

À l'étude de l'ensemble de paires rapprochées, une autre originalité retient notre attention : la conservation d'un trait / concret / par les mots *autoroute* et *navigation*, qui le portent dans leur sens original mais ne sont employés dans le sous-corpus « nouvelles technologies » que d'une manière métaphorique totalement codifiée (« autoroutes de l'information », « navigation sur Internet »). Cela est probablement imputable à un choix de la part des auteurs des textes étudiés consistant à se placer ou non sur le registre d'un parallèle physique au « cyberespace », et à s'y maintenir de manière cohérente.

[9]Dans le cas de la paire *dimension* / *échelle*, c'est plus rigoureusement ce dont on évalue la dimension ou l'échelle qui peut être qualifié de matériel ou immatériel.

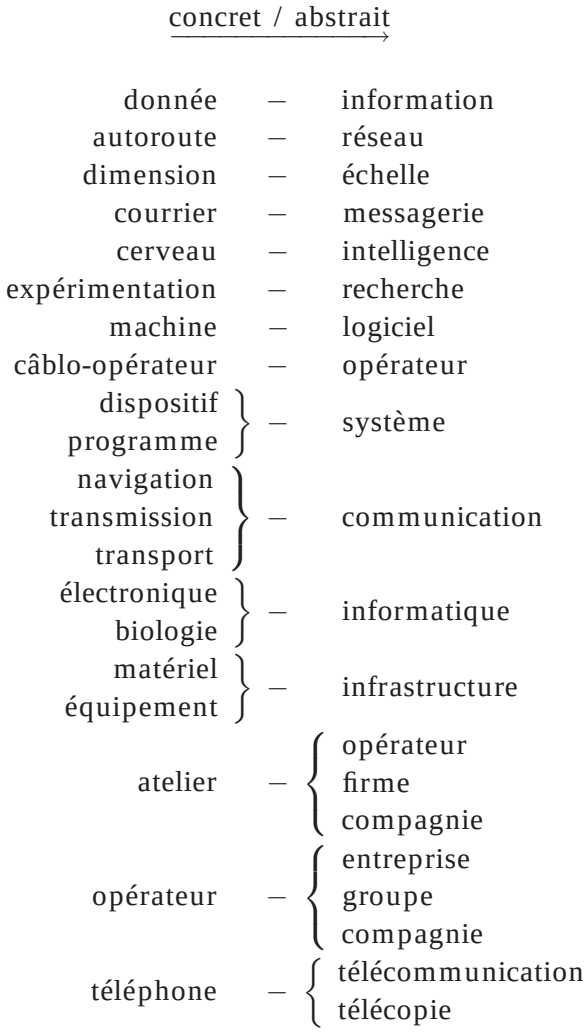


FIG. 4.6 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / concret - abstrait /

ancien - nouveau Les paires dont les termes sont distingués selon une évaluation temporelle de modernité — que nous désignons par le sème spécifique / ancien - nouveau / — sont présentées sur la figure 4.7.

Nous ne nous étendons pas cette fois sur la sémantique de la relation exposée, mais apportons en revanche une précision concernant l'opposition retenue entre *firme* et *corporation*, d'une part, et *entreprise*, *compagnie* et *groupe*, d'autre part. Bien que cette alternance semble justifiée par l'usage courant (la connotation de *firme* étant assez « rétro », et celle de *corporation* tout à fait Ancien Régime), nous avons pu constater par un retour aux données textuelles que l'apparition de *corporation* dans cette étude était en particulier due à une erreur d'analyse morphosyntaxique de noms de compagnies améri-

ancien / nouveau →

technique	–	technologie
électronique	–	informatique
câble	–	satellite
radio	–	télévision
téléphone, téléphonie	–	télécommunication
firme, corporation	–	entreprise, groupe, opérateur

FIG. 4.7 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / ancien - nouveau /

simple / sophistiqué →

technique	–	technologie
électronique	–	informatique
câble	–	satellite
téléphonie	–	télécommunication
dispositif	–	programme
radio	–	télévision

FIG. 4.8 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / simple - sophistiqué /

caines, en particulier de la compagnie de téléphonie Bell Corporation. Le mot *corporation* se trouve ainsi influencé par son association avec la technologie « dépassée » qu'est le téléphone.

simple - sophistiqué La figure 4.8 présente les paires de mots distinguées par un degré de sophistication supérieure attaché à leur second terme.

De manière peu surprenante, il est possible d'observer de nombreux points communs entre les paires précisées par / ancien - nouveau / et / simple - sophistiqué /.

condition La distinction que nous nommons / condition /, présentée sur la figure 4.9, recouvre en réalité plusieurs nuances de sens plus fines ; c'est pourquoi des lignes pointillées découpent sur cette figure en trois sous-ensembles les paires de mots considérées.

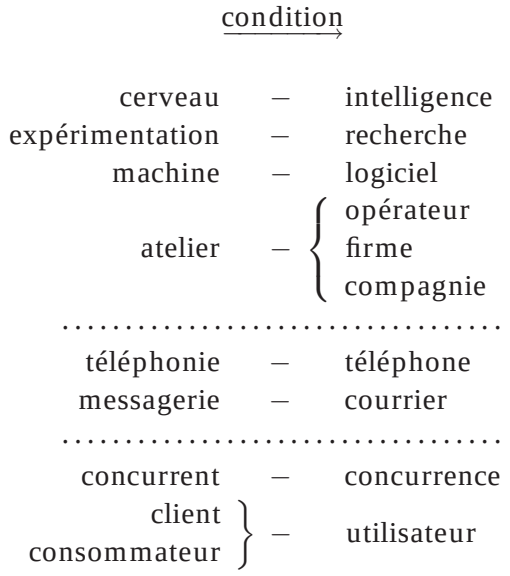


FIG. 4.9 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / condition /

Le premier groupe rassemble les paires de mots simultanément distingués par les sèmes spécifiques / condition / et / concret - abstrait / ; dans ce cas, le premier terme constitue la condition matérielle, le « substrat » permettant le développement de l'abstraction qu'est le second terme. Les paires du second groupe, au contraire, sont dans une relation / abstrait - concret / ; on voit alors apparaître la technologie (immatérielle) sous-jacente et nécessaire à l'existence d'une « commodité » matérielle.

Les paires du troisième groupe ne sont précisées par aucune autre relation, et leur interprétation est donc plus libre. La distinction entre *concurrent* et *concurrence* pourrait être rattachée au premier groupe mentionné ci-dessus, mais notre analyse ne permet pas de faire apparaître entre ces mots la relation / concret - abstrait /. Celle qui oppose *client* et *consommateur* à *utilisateur* peut paraître plus discutable, et seule notre connaissance de l'arrière plan philosophique marxiste de nombreux contributeurs du *Monde diplomatique* a pu nous amener à retenir cette association comme valide : l'un des lieux communs de l'analyse marxiste du système capitaliste est en effet la nécessité pour un produit de consommation de réaliser sa valeur marchande avant de pouvoir exprimer sa valeur d'usage — la première ayant un effet d'« écran » sur la seconde. L'opposition lexicale que nous avons mise au jour n'est qu'une traduction de ce principe du point de vue du consommateur : l'accession au statut d'utilisateur est conditionnée par un passage par le statut de client.

texte / image
⟶

imprimerie	–	photographie
tirage	–	audience
télécopieur	–	téléviseur
copie	–	cassette

FIG. 4.10 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / texte - image /

texte - image Les paires rassemblées à la figure 4.10 n'appellent que peu de commentaire, à l'exception peut-être de *copie* / *cassette*, pour laquelle la distinction / texte - image / n'est valide que dans un sens de ces mots (*copie* comme exemplaire d'un livre, *cassette* comme « cassette vidéo »), sens dont la prédominance n'est pas nécessairement prescrite par le domaine « nouvelles technologies ». Nous avons néanmoins fait le choix de l'inclure en raison de l'importance statistique de ses liens avec les autres paires de la classe.

production - consommation La figure 4.11 présente les paires de mots distinguées par le sème spécifique / production - consommation /, qui n'appelle lui non plus aucune remarque particulière.

production / consommation
⟶

présentateur	–	lecteur
présentation	–	lecture
éditeur	–	lecteur
rédacteur	–	éditeur

FIG. 4.11 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / production - consommation /

agressif - consensuel Nous terminons cette présentation par la distinction dont la validité est la plus conditionnée par le corpus étudié, que nous avons désignée par / agressif - consensuel / (*cf.* figure 4.12).

Remarquons tout d'abord que beaucoup des mots distingués par ce sème spécifique ne sont pas centraux au domaine « nouvelles technologies », mais relèvent plutôt de l'analyse des mécanismes des marchés financiers — ces deux thèmes étant souvent liés dans *Le Monde diplomatique*. La paire « prototypique » de cette relation est *rachat* / *fusion*, dont les termes correspondent

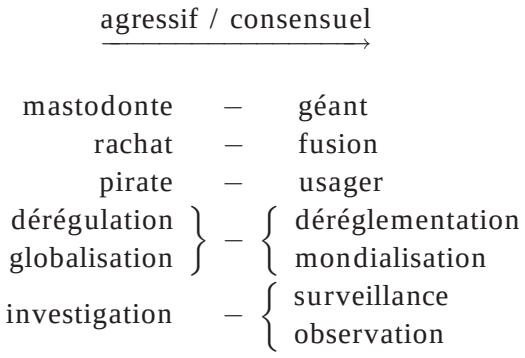


FIG. 4.12 – *Paires de mots du domaine « nouvelles technologies » distinguées par le sème spécifique* / agressif - consensuel /

sur les marchés financiers à des degrés de conciliation nettement distincts. S'y associe l'alternative *mastodonte* / *géant*, dont le second terme est relativement neutre et le premier beaucoup plus menaçant — imagine-t-on le roman de Roald Dahl *Le Bon Gros Géant* rebaptisé *Le Bon Gros Mastodonte* ? En ce qui concerne la paire *dérégulation* / *déréglementation*, le second de ces mots désigne la disparition des règlements, et le premier celle de toutes les règles — on est donc bien dans une forme de capitalisme plus agressive. La nuance est similaire entre *globalisation* et *mondialisation*, mais il s'agit là plus d'une « convention tacite d'usage » parmi les contributeurs du *Monde diplomatique* que d'un sens réellement codifié.

De manière plus centrale au domaine considéré, nous avons retenu ce même sème spécifique comme valide pour distinguer *investigation* de *surveillance* ou *observation*, le premier de ces mots correspondant à une attitude plus active, plus intrusive. La distinction entre *pirate* et *usager* le long de cet axe est pour sa part évidente, et illustre tout l'intérêt interprétatif de la méthode de recherche de sèmes spécifiques par « analogie de distinction » que nous avons développée : des « voyous du capitalisme » (*mastodonte*) à ceux de l'informatique (*pirate*), une même nuance sémantique subsiste…

Ayant présenté individuellement les divers sèmes spécifiques mis au jour par notre analyse, nous proposons à la section suivante une vue générale de la structuration du domaine « nouvelles technologies » qu'ils induisent conjointement.

4.5.4. Vue d'ensemble des informations acquises

La figure 4.13 rassemble les informations acquises concernant la structuration d'un sous-ensemble des taxèmes du domaine « nouvelles technologies » présentés à la figure 4.5, en adoptant une représentation des relations entre mots sous forme de réseau sémantique parfois employée par F. Rastier dans

ses ouvrages, et reprise également dans [PS99]. Le choix de ne considérer qu'un sous ensemble des taxèmes structurés résulte d'un souci de clarté, et leur sélection des taxèmes est guidée par l'intérêt des structurations exposées par la représentation graphique ; nous n'avons en particulier pas fait apparaître sur la figure les nombreux taxèmes réduits à deux mots distingués par un sème spécifique simple.

Une première constatation méritant d'être exprimée à l'étude de cette représentation est que, si la structuration en « sous-classes » sémantiques des taxèmes telle que réalisée par les travaux décrits à la section 4.2 [Den04, Cou05] pouvait permettre de faire apparaître des regroupements suggérant des sèmes spécifiques, la réciproque est également vraie. Dans les classes de tailles relativement importantes, les distinctions entre mots amènent certains d'entre eux à se regrouper pour former des « noyaux » de termes plus proches de la synonymie. Ainsi, dans le taxème {*atelier, câblo-opérateur, opérateur, firme, corporation, entreprise, compagnie, groupe*}, les oppositions similaires distinguant *compagnie, groupe* et *entreprise* des autres mots les conduisent à se rassembler en un « noyau dense ».

Il est également intéressant de remarquer que les sèmes spécifiques mis au jour permettent de distinguer les sémèmes au sein des taxèmes, mais relèvent également du niveau du discours. En particulier, le fait que plusieurs paires de sémèmes puissent être simultanément distinguées selon plusieurs axes suggère l'existence d'un faisceau* d'isotopies « / ancien-moderne / + / concret-abstrait / + / simple-sophistiqué / », très révélateur du discours intellectuel assez courant au début de la période de la « révolution des technologies de l'information », avec laquelle coïncident les dates des archives du *Monde diplomatique* que nous avons étudiées.

Cette convergence entre structuration microsémantique du lexique et organisation du discours n'est certes pas surprenante si l'on considère l'échelle textuelle que nous avons exploitée pour faire apparaître les sèmes spécifiques. Il n'est pourtant pas évident *a priori* que la langue et le lexique puisse trouver des points de contacts aussi importants, et les résultats obtenus sont donc particulièrement intéressants en tant que confirmation de ce fait.

4.5.5. Vers une représentation explicite des sèmes spécifiques

Le principal défaut d'automatisation persistant dans la méthode proposée est la nécessité de recourir au jugement humain pour désigner les sèmes spécifiques mis au jour. Nous n'avons pu au cours de nos expériences définir de procédure permettant d'informatiser ne serait-ce que partiellement cette tâche, par exemple en recherchant les indices les plus discriminants justifiant les regroupements de paires de mots finalement validés. De nombreuses pistes restent néanmoins à explorer afin de résoudre cette difficulté ; on peut par exemple envisager de réaliser à l'issue du processus de validation des classes

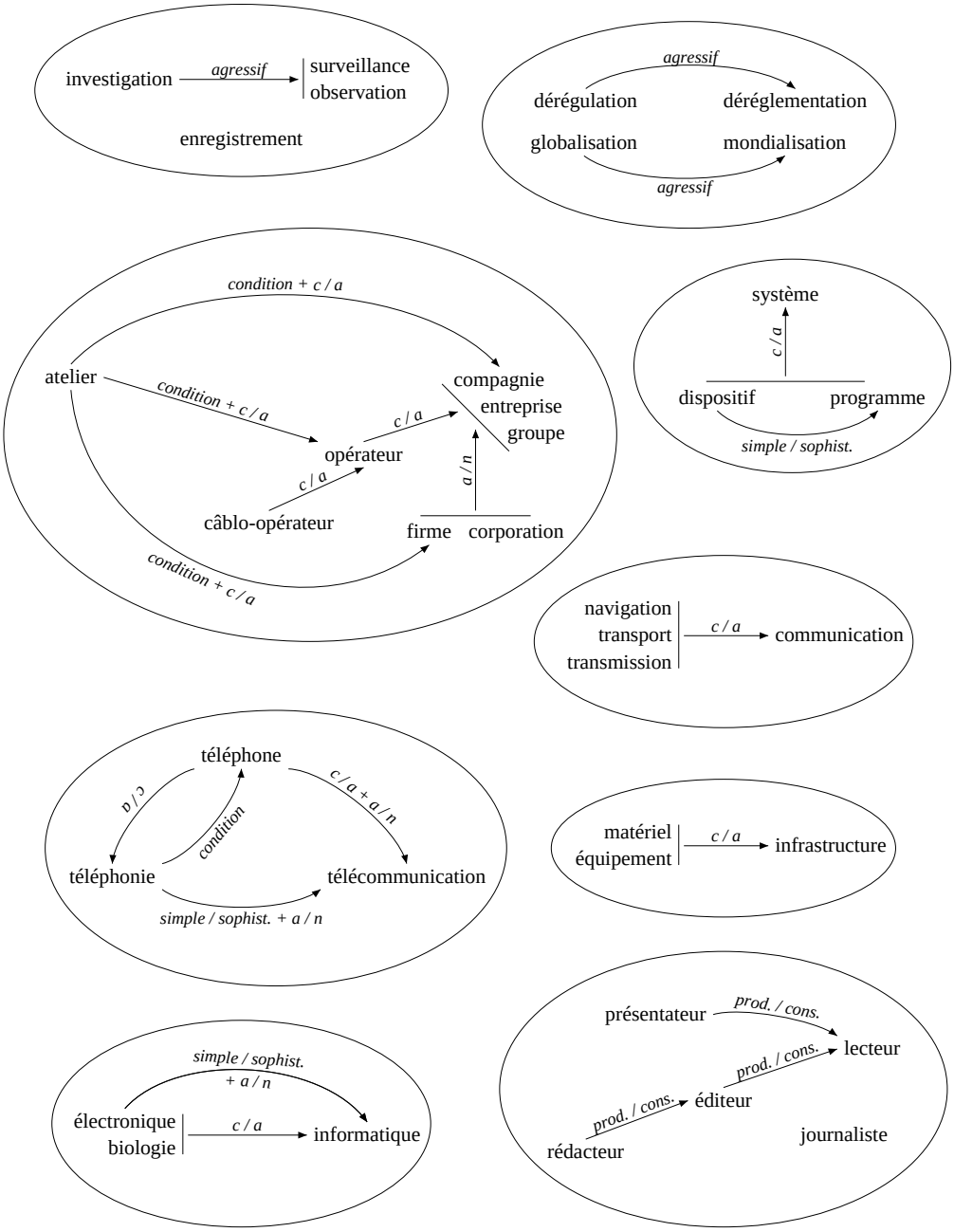


FIG. 4.13 – *Structuration d'une sélection de taxèmes du domaine « nouvelles technologies » par des sèmes spécifiques. Une flèche reliant deux mots indique que leur sens est distingué par le sème porté par la flèche. Les abbréviations utilisées sont « a/n » pour « ancien/nouveau », « c/a » pour « concret/abstrait », « prod./cons. » pour « producteur/consommateur » et « simple/sophist. » pour « simple/sophistiqué »*

une nouvelle sélection d'indices contextuels, selon des critères distincts, au sein desquels une verbalisation des sèmes spécifiques serait plus probable.

Toutes les expériences que nous avons menées dans ce chapitre s'appuient sur les données textuelles collectées pour le domaine « nouvelles technologies ». Nous évaluons dans la section suivante la qualité des informations qu'il est possible d'extraire pour d'autres thèmes.

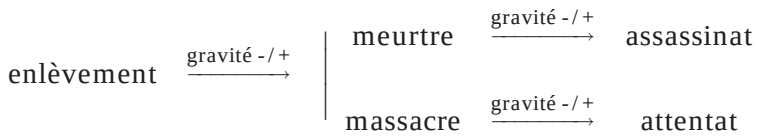
4.6. Maintien de la qualité des résultats sur d'autres domaines

Outre des facteurs « évidents » tels que la quantité de données textuelles disponible pour l'analyse ou la qualité des classes sémantiques étudiées, la méthode proposée afin de mettre au jour des sèmes spécifiques au sein des taxèmes d'un domaine est tributaire du type de discours développé autour d'un thème donné. En effet, elle consiste pour ainsi dire à replonger dans le lexique l'organisation du discours, approche d'autant plus efficace que les textes étudiés présentent un discours « stéréotypé », faisant couramment usage des mêmes idées et oppositions dialectiques.

En conséquence, on observe des résultats assez peu concluants dans le domaine < dépense / réduction / budget >, alors que celui-ci correspond à un thème largement abordé dans *le Monde diplomatique*. Cela s'explique par le fait que les préoccupations budgétaires dont il est question peuvent être évoquées dans des contextes très divers (du porte-monnaie individuel au budget d'un état), en employant suivant les auteurs des outils d'analyse différents, avec des opinions potentiellement divergentes.

Ce type de situation constitue néanmoins une exception, la plupart des domaines sur lesquels nous avons réalisé une étude se prêtant à une structuration plus ou moins élaborée. En moyenne quatre ou cinq sèmes spécifiques (de deux à six dans les cas extrêmes), d'importance variable, peuvent être mis au jour. Ils permettent selon les domaines de structurer le contenu de 20 à 60 % des classes sémantiques étudiées — ce pourcentage étant légèrement supérieur si l'on dénombre les mots ainsi caractérisés, car les classes les plus fournies sont celles qui donnent le mieux prise à l'analyse. Même si la méthode développée ne permet en général pas de faire apparaître tous les sèmes spécifiques « majeurs » qu'une étude manuelle des taxèmes du domaine pourrait faire espérer, elle peut en revanche mettre au jour des distinctions plus inattendues. Nous présentons ici à titre d'illustration quelques exemples de résultats obtenus pour les domaines < violence / victime / crime >, < film / art / scène > et < télévision / chaîne / publicité >. La méthodologie adoptée pour la validation des résultats est identique à celle que nous avons mise au point pour l'analyse du domaine « nouvelles technologies » ; comme lors de la présentation des résultats de cette première expérience, les rapprochements entre paires de mots sont suggérés de manière automatique, la sélection des regroupements les plus pertinents et leur interprétation restant manuelle.

Domaine < **violence / victime / crime** > La principale structuration observable dans les taxèmes de ce domaine est celle d'une gradation de la gravité des faits relatés, que nous notons / gravité -/+ /. La paire « prototypique » de ce sème spécifique est « délit / crime » (au sein du taxème {*crime*, *délit*, *infraction*}). Cette même relation permet de structurer le taxème {*enlèvement*, *massacre*, *attentat*, *meurtre*, *assassinat* } par :



En revanche, l'aspect prémédité de *assassinat* et *attentat*, qui justifie leur qualification de gravité supérieure par rapport à *meurtre* et *massacre*, respectivement, n'est pas discernable dans les associations proposées. Deux autres taxèmes voient leur contenu spécifié par le sème / gravité -/+ / :

Au sein du taxème {*enlèvement*, *massacre*, *attentat*, *meurtre*, *assassinat*}, un sème spécifique permet de distinguer *meurtre* de *massacre* ainsi que *assassinat* de *attentat* : nous supposons que la distinction se fait ici selon le nombre de victimes (unique / multiples).

De manière moins centrale au thème traité, on peut également remarquer dans les taxèmes {*accusation*, *dénonciation*, *témoignage*} et {*procès*, *jugement*} une distinction identique entre *dénonciation* et *accusation*, *dénonciation* et *témoignage*, ainsi que *jugement* et *procès* : la nuance est ici marquée entre justice « informelle » (un jugement n'étant pas nécessairement le fait d'un tribunal) et « procédurale ».

Enfin, le taxème {*père*, *mère*, *fils*, *fille*} donne lieu à la mise au jour de deux sèmes spécifiques, l'un distinguant selon le sexe *père* de *mère* et *fils* de *fille*, et l'autre selon la génération *père* de *fils* et *mère* de *fille*.

D'autres sèmes spécifiques dont on aurait pu attendre l'apparition, comme la distinction entre civils et militaires parmi les acteurs des événements relatés, ou entre forces légales (*police*, *armée*, *etc.*) et criminelles (*gang*, *mafia*, *etc.*) dans les taxèmes regroupant des groupes organisés, ne sont en revanche pas mis au jour par le système.

Domaine < **film / art / scène** > Nous ne détaillons pour ce domaine que sa structuration la plus spécifique et notable. Elle présente la particularité d'être réalisée par des sèmes spécifiques ne s'exprimant pas de manière privilégiée

sous la forme d'une gradation ou opposition, mais prenant la forme de « sous-domaines ». Ainsi *dramaturge* se distingue de *cinéaste* comme *comédien* de *acteur* (théâtre / cinéma), *chanteur* s'oppose à *écrivain* comme *musicien* à *comédien* (musique / texte), *etc.* Ces relations peuvent s'exprimer de manière plus intuitive sous forme d'un tableau répartissant les mots des trois principaux taxèmes concernés selon leurs sèmes spécifiques :

Classe	/ théâtre /	/ cinéma /	/ texte /	/ musique /
{*public*, *téléspectateur*, *spectateur*, *lecteur*}	*spectateur*	*téléspectateur*	*lecteur*	-
{*intellectuel*, *artiste*, *écrivain*, *cinéaste*, *romancier*, *peintre*, *poète*, *chanteur*, *dramaturge*}	*dramaturge*	*cinéaste*	*écrivain* *poète*	*chanteur*
{*acteur*, *comédien*, *musicien*}	*comédien*	*acteur*	*comédien*	*musicien*

La répartition n'est pas aussi « parfaite » que l'on pourrait le souhaiter : le système n'a ainsi pas permis de caractériser de manière pertinente *romancier*, et la relation entre / théâtre / et / texte / est assez délicate à préciser — le second semble d'après certains des regroupements effectués « généraliser » le premier, mais pas de manière systématique. Une opposition n'est pas représentée sur le tableau, car elle s'exprime pour sa part en terme réellement différentiels : *artiste* se distingue de nombreux mots de la deuxième classe présentée selon une nuance détachant également *public* des autres membres de la première classe. Après le cas des interprétations « limites » du sème / concret - abstrait / dans le domaine « nouvelles technologies », nous sommes ici une nouvelle fois en présence d'un sème spécifique exprimant une relation d'hyperonymie.

Domaine < télévision / chaîne / publicité > Nous concluons cet aperçu rapide des résultats observables dans divers domaines lexicaux par l'étude de < télévision / chaîne / publicité >.

Une première nuance remarquable dans ce domaine révèle la présence ou absence d'une référence financière explicite dans un mot : ce sème spécifique permet de distinguer *recette* et *vente* de *audience*, *producteur* de *média*, et enfin *client* et *consommateur* de *public*. On peut également noter la présence d'une alternance / simple - multiple / dans les paires *chaîne* / *réseau*, *film* / *feuilleton*, et *film* / *série*.

Mais surtout, les relations mises au jour dans ce domaine proposent enfin une distinction « sensée » permettant d'identifier *client* et *consommateur*, paire ayant fait l'objet de notre attention à la section 4.4.2. Le mot *client* se trouve, selon les rapprochements opérés par le système développé, distingué de *consommateur* par une nuance similaire à celle séparant *marketing* de *promotion*, tout comme *propagande* de *marketing*. Se dessine donc un parallèle

entre une sophistication croissante des techniques commerciales (*propagande* dépassant même ce cadre), d'une part, et de la relation commerciale (rejoignant l'observation faite à la section 4.4.2 de l'expression « conversation avec le client »), d'autre part. Il est assez délicat d'exprimer simplement de cette gradation ; nous pouvons par exemple la désigner, tout imparfaitement que cela soit, par / sophistication commerciale -/+ / ou, dans une interprétation légèrement différente, / psychologie -/+ /.

4.7. Conclusion

Nous avons présenté dans ce chapitre un premier système permettant de mettre au jour, de manière semi-automatique et à partir des seules données textuelles, des éléments de distinction sémantique fine entre mots appartenant à une même classe, atteignant ainsi un degré de définition du sens encore inédit en acquisition automatique sur corpus d'information sémantiques lexicales. Conformément au choix théorique qui guide l'ensemble de notre étude, ces éléments de distinction sont formalisés comme sèmes spécifiques de la sémantique différentielle, et la méthode que nous avons développée pour les mettre au jour s'inspire également des principes de cette théorie. La technique mise au point réalise en s'appuyant sur l'étude de grands voisinages autour des occurrences des mots comparés un regroupement de paires de mots distingués par des nuances similaires. Cette approche présente, par rapport à des travaux antérieurs [PS99], la faiblesse de ne permettre qu'une représentation implicite des sèmes découverts : les listes de paires de mots rapprochées doivent être interprétées afin de « comprendre » l'élément de distinction commun. En revanche, elle permet une automatisation bien plus importante du processus de recherche de sèmes spécifiques, suffisante pour assurer que le travail humain de validation des résultats ne puisse introduire dans ceux-ci de biais interprétatif trop important.

Un autre intérêt de la représentation implicite des sèmes spécifiques par rapport à leur représentation explicite est le degré de généricité qu'elle permet d'atteindre. En effet, la recherche automatique de désignations explicites de sèmes spécifiques parmi les voisinages de mots comparés laisse ouverte la possibilité qu'une même distinction soit désignée, par exemple, une fois par « grand », une autre par « gros », une troisième par « large », *etc.*, sans qu'il soit *a priori* possible de détecter que c'est dans tous ces cas la même nuance qui est en jeu. Si l'apposition automatique d'une « étiquette » désignant le sème spécifique est plus « parlante » pour un utilisateur humain, elle est donc en revanche bien moins intéressante en ce qui concerne leur possibilité d'exploitation interprétative : la variété des dénominations possibles d'un même sème rend apparemment impossible la détection d'isotopies sémantiques spécifiques. À l'inverse, la méthode que nous proposons est à même de rapprocher jusqu'à 20 paires de mots distingués par une même nuance, ce qui nous permet d'es-

pérer pouvoir mettre au jour des structurations de haut niveau du discours par la détection d'isotopies spécifiques. Reste maintenant afin d'obtenir un outil interprétatif complet à exposer les liens existant entre sèmes spécifiques et génériques, afin de pouvoir détecter et exploiter les occurrences d'isotopies « mixtes ».

Cette problématique, tout comme de nombreuses autres soulevées par nos recherches, reste ouverte à l'issue de notre étude. Nous réalisons au chapitre suivant, en conclusion de ce mémoire, un bilan des apports de nos travaux, et soulignons les pistes qui nous paraissent les plus prometteuses pour des développements à venir.

Conclusion

Nous avons présenté dans ce document une méthodologie générique d'acquisition automatisée sur corpus d'informations sémantiques lexicales fondées sur la sémantique différentielle. Elle s'organise en trois étapes calquées sur les principes de la théorie linguistique adoptée, au cours desquelles nous avons tenté de mener aussi loin que possible les techniques d'analyse statistique de données qui constituaient notre base technique.

Dans un premier temps, nous avons réussi à regrouper les mots du lexique en *domaines*, notion que nous avons approximée par celle de « vocabulaire particulièrement associé à l'évocation d'un thème ». Nous avons développé afin de résoudre ce premier problème le système FAESTOS, qui construit par analyse des répartitions des mots du corpus sur ses paragraphes des classes de mots-clés permettant tout à la fois de caractériser les principaux thèmes abordés dans les textes étudiés et de détecter leurs occurrences. Ces classes sont en outre d'une qualité suffisante pour permettre une « vue d'ensemble interprétative » du corpus étudié, de son contenu et du type de discours qu'il développe. FAESTOS fonctionne de manière totalement autonome, sans requérir de donnée auxiliaire ou d'intervention manuelle. Cela est rendu possible par la mise au point d'algorithmes originaux d'analyse statistique de données, permettant en particulier de répartir des objets en classes à partir d'une classification hiérarchique, en remettant éventuellement en cause les agrégations proposées par celle-ci et sans recourir à aucun présupposé concernant le nombre ou la taille de ces classes, par confrontation de deux classification hiérarchiques « réciproques ». Notons toutefois que si les classes de mots-clés constituées par le système sont à la fois « sensées » et cohérentes, leur performance pour la détection de thèmes n'atteint qu'un rappel assez modéré (de l'ordre de 65 %). À l'issue de l'ensemble de traitements décrits au chapitre 2, le vocabulaire étudié est réparti dans des ensembles domanialement cohérents, auxquels sont associés des sous-corpus rassemblant les paragraphes évoquant les thèmes correspondant.

La seconde étape de construction de lexiques a pour but de rassembler les mots de chacun de ces domaines en *taxèmes*, ou classes sémantiques de mots de sens proches, permutables en contexte. Il est à cette fin fait usage de mesures de similarité évaluant la ressemblance entre les manières dont les mots sont employés dans les données textuelles, à partir des données de leurs

contextes d'apparition immédiats. La solution que nous proposons consiste à réaliser la construction de classes sémantiques en deux temps : tout d'abord, une méthode ensembliste relativement simple est employée pour calculer une première approximation des similarités d'usage entre mots en exploitant la totalité des données textuelles rassemblées dans le corpus. Cette première connaissance est ensuite réinvestie dans le calcul des similarités sémantiques entre mots au sein des domaines, n'exploitant que les données textuelles présentes dans le sous-corpus thématique associé. L'intérêt de cette approche est en particulier de contourner les difficultés induites par la taille restreinte de ces sous-corpus. Ces recherches ont été l'occasion de la mise au point de deux techniques novatrices d'analyse statistique de données permettant, pour la première, la normalisation *a posteriori* d'un ensemble de mesures de similarité, et pour la seconde, le calcul de similarités normalisées entre mots par échantillonnage des voisinages caractérisant leurs usages. En proposant avec cette seconde méthode une technique générique d'évaluation des similarités entre mots à partir des similarités calculées entre leurs contextes considérés individuellement, nous permettons une séparation nette entre les aspects statistique et proprement linguistique de la question, simplifiant ainsi le problème étudié de « comment comparer deux mots à partir de l'ensemble de leurs usages constatés » en « comment évaluer la similarité entre deux voisinages de mots ». Nous espérons que cette redéfinition du problème pourra être l'occasion de nouveaux progrès dans le domaine de la construction de classes sémantiques. Les mesures de similarité sémantiques entre mots présentées au chapitre 3 nous permettent de construire par une méthode de classification ascendante hiérarchique classique des arbres de classification de mots de relativement bonne qualité, quoique les regroupement proposés entre mots restent de pertinence parfois inégale, et puissent être « parasités » par des mots moins immédiatement similaires. Leur exploitation afin d'en extraire des classes sémantiques, ou taxèmes, reste, dans l'état actuel de nos recherches, toujours manuelle.

La troisième étape, enfin, est consacrée à la recherche au sein des classes ainsi définies de *sèmes spécifiques*, ou traits distinctifs permettant d'individualiser les sens de mots dans un taxème. Il s'agit là d'un domaine de recherche encore quasi vierge, dans lequel nous n'avons pu réaliser qu'une première incursion — en particulier, notre étude se limite au seul cas des noms. Nous proposons une méthodologie assez simple mais originale permettant de regrouper des paires de mots proches apparemment distingués par les mêmes nuances, réalisant ainsi une caractérisation « implicite » des distinctions de sens entre ces mots. Cette technique ne peut néanmoins être considérée comme une réelle automatisation, mais plutôt comme un outil de proposition, supposant encore un travail de « tri » manuel relativement conséquent afin d'aboutir à des résultats qui fassent sens. Nonobstant cette limitation, nous montrons qu'il est possible de structurer semi-automatiquement les classes sémantiques d'un domaine de manière cohérente, avec une finesse de définition de sens encore ja-

mais atteinte à notre connaissance dans le domaine de l'acquisition sur corpus d'informations lexicales sémantiques.

Dans leur ensemble, les travaux que nous avons menés sont riches d'enseignements encourageants, ne serait-ce que d'avoir confirmé que la possibilité d'acquisition d'informations lexicales sémantiques, même très fines, s'étend au cas de corpus de langue complexe et peu codifiée : en effet, même les corpus dits « généralistes » classiquement employés dans ce domaine (typiquement, *Le Monde* pour le français et le *Wall Street Journal* pour l'anglais) relèvent généralement d'une langue journalistique assez directe et factuelle, alors que *Le Monde diplomatique* rassemble pour sa part des constributions plus analytiques d'auteurs tels que Guy Debord, Cornelius Castoriadis ou Jacques Derrida. Plus concrètement, outre les avancées techniques et méthodologiques déjà mentionnées plus haut lors de notre retour sur les trois étapes du travail effectué, deux apports principaux retiennent notre attention.

Tout d'abord, notre recherche a permis l'identification de paliers textuels pour l'analyse adaptés aux types de relations paradigmatiques recherchées. Les choix de faire usage de paragraphes entiers pour la recherche de thèmes (échelle adaptée, comme nous l'avons constaté, au développement d'isotopies domaniales) et de voisinages immédiats pour la construction de classes sémantiques (selon un principe de cohésion textuelle à courte distance) sont relativement « classiques » et consensuels. La nécessité de remonter à un palier textuel supérieur, celui du discours argumentatif, pour mettre au jour les sèmes spécifiques peut sembler plus surprenante. Les sèmes spécifiques constituent en effet selon la sémantique différentielle les « plus petits » éléments de définition du sens, et une approche logique élémentaire (ou simpliste) suggèrerait pour les faire apparaître de concentrer l'étude sur une échelle textuelle plus petite que pour la constitution de taxèmes. Nos expériences tant théoriques que pratiques tendent à montrer que ce n'est pas le cas, et que les sèmes spécifiques interviennent plus volontiers dans le processus interprétatif à l'échelle dialectique. Cela constitue un premier élément de connaissance important pour l'étude de ces composants sémantiques encore mal connus.

D'un point de vue plus technique, les résultats que les méthodes développées permettent d'obtenir tendent à conforter notre choix de ne faire appel à l'analyse syntaxique à aucune étape des recherches menées (suivant en cela la défiance exprimée par F. Rastier à l'égard des approches de l'interprétation guidées par la syntaxe). À l'évidence, les relations grammaticales couramment employées dans le domaine de l'acquisition d'information lexicales (comme la relation tête-modifieur) ne constituent pas un outil pertinent pour l'analyse de phénomènes textuels de l'échelle du paragraphe (recherche de domaines), ou même de la phrase considérée dans son ensemble (recherche de sèmes spécifiques). En ce qui concerne la construction de classes sémantiques, toutefois, la quasi totalité des travaux existants fait usage d'outils permettant d'extraire

les mots en lien syntaxique direct avec un mot à caractériser. Les exemples que nous présentons tendent à prouver qu'une approche faisant simplement usage de la notion de proximité permet d'extraire une quantité importante de cette information. Nous montrons en outre que ces mots directement liés par la syntaxe ne sont pas les seuls à participer à la contrainte syntagmatique restreignant l'ensemble des mots utilisables à un endroit donné dans un énoncé — comme dans l'exemple déjà mentionné de l'expression « catégorie syntaxique de mot », où « syntaxique » constitue un indice intéressant pour préciser le sens de « mot », même s'il n'est pas en lien direct avec ce dernier.

Ayant brièvement rappelé les apports et enseignements de nos travaux, nous souhaitons dans ce qui suit définir, parmi les pistes de recherche qu'ils suscitent, celles qui nous paraissent les plus prometteuses. Nous ne revenons pas ici sur les évolutions possibles du système ARTEMIS, déjà présentées au chapitre 2, qui constituent pour la plupart, plus que de réels axes de recherche, des évolutions techniques de ce qui est déjà un système à part entière.

En ce qui concerne la tâche de construction de taxèmes, nous avons déjà évoqué la nécessité d'un approfondissement plus proprement linguistique de la notion de ressemblance sémantique entre deux énoncés (ou voisinages). Ce travail, tout nécessaire qu'il soit, ne peut néanmoins sans doute pas suffire à surmonter la nécessité d'une extraction manuelle de classes sémantiques depuis la classification obtenue. Une constatation s'impose pourtant : si cette seconde étape de l'analyse paraît moins aboutie que la première, l'arbre de classification sémantique des mots construit au chapitre 3 est de qualité au moins équivalente aux premiers arbres de classification thématique construits au chapitre 2. C'est donc moins la qualité de cette classification elle-même qui est source d'insatisfaction, que l'absence d'un second critère d'évaluation qui nous permettrait, comme lors de la construction des classes de mots-clés thématiques, d'effectuer une lecture « intelligente » de l'arbre de classification. Nous n'avons pu au cours de nos recherches trouver un tel critère au sein des textes eux-mêmes. En revanche, une solution à ce problème se présente si nous disposons pour réaliser notre étude d'un corpus parallèle multilingue, présentant le même contenu sémantique sous des formes distinctes. En combinant les tâches d'alignement de corpus et de construction de classes sémantiques pour chacune des langues en présence, il est envisageable à la fois d'améliorer les performances de l'alignement (le fait de considérer des classes sémantiques permettant de contourner les difficultés d'alignement liées à la non identité de traduction des diverses occurrences d'un mot) et celles de la construction des classes sémantiques, la contrainte de similarité de localisation dans les textes des mots de classes correspondantes en différentes langues constituant le « critère externe » qui nous manque pour réaliser une analyse critique de la classification sémantique.

L'extraction de sèmes spécifiques, pour sa part, n'en est qu'à ses tout débuts, et nous ne prétendons pas baliser à la seule lumières de nos expériences toute la potentialité des approches possibles pour leur extraction. Une première nécessité est naturellement d'étendre les recherches aux autres catégories morphosyntaxiques « pleines » que le nom, sur lesquels se sont focalisés nos efforts. Les études préliminaires que nous avons réalisées dans ce sens suggèrent que, le phénomène linguistique considéré étant plus « fin » que les domaines et taxèmes, des techniques différentes devront sans doute être développées et adaptées à chaque catégorie de mots. Loin d'être un handicap, cela constitue au contraire une source de richesse potentielle très importante : si par exemple un travail sur les adjectifs permet l'identification au sein de ceux-ci d'une opposition « mélioratif / dépréciatif » non identifiée par ailleurs pour les noms, il peut être possible de la projeter sur ceux-ci grâce aux associations noms-adjectifs.

Plus essentiellement, nous avons identifié le principe d'une détermination des sèmes spécifiques par analyse des phrases d'apparition des mots à caractériser considérées dans leur totalité : c'est là à la fois trop et trop peu d'information… Trop car, naturellement, s'il existe dans la phrase considérée une isotopie à laquelle prend part un sème spécifique du mot, celle-ci n'impliquera qu'une faible proportion des mots de la phrase ; il est donc nécessaire dans un premier temps d'étudier empiriquement les configurations de ces isotopies afin de tenter d'orienter plus efficacement la sélection des indices, l'« éparpillement » de ceux-ci justifiant d'ailleurs peut-être que l'approche purement statistique que nous avons adoptée soit abandonnée en faveur de techniques d'apprentissage symbolique. Trop peu, car le choix de l'échelle textuelle considérée elle-même demande également à être approfondi : si nous restons persuadé que le niveau du développement des idées est bien le plus pertinent pour la tâche entreprise, l'approximation de celui-ci par la phrase est un pis-aller discutable — la définition par F. Rastier d'une isotopie sémantique indiquant par exemple clairement son caractère potentiellement transphrastique. Une fois encore, il paraît nécessaire afin de découvrir les indices pouvant permettre un « découpage » optimal de soumettre les données disponibles à une exploration linguistique rigoureuse, seule approche susceptible d'élever l'analyse informatique de données textuelles au-dessus de la simple manipulation de chaînes de caractères.

Annexes

Annexe A

Résultat de FAESTOS, système d'extraction de classes de mots-clés thématiques

Nous présentons ici les 46 classes de mots-clés thématiques obtenues par une exécution du système FAESTOS exploitant les données rassemblées dans le corpus d'archives du *Monde diplomatique* (les 5 classes « anecdotiques » automatiquement écartées par le système sont également reportées ici, en fin de liste, une note permettant de les identifier). Le recours effectué par le système à une procédure de sélection aléatoire d'extraits du corpus étudié implique de très légères variations des résultats d'une exécution du système à l'autre, ce qui explique que les classes présentées ici ne correspondent pas exactement à celles dont nous avons présenté les dénominations à la section 2.8.

Pour chaque classe, nous indiquons le nom qui lui a été automatiquement attribué par le système, le nombre de mots qu'elle rassemble, le nombre de paragraphes du corpus où elle détecte une occurrence de « son » thème, puis la liste des mots qui la composent. Les figures A.1 et A.2, page suivante, proposent en outre une vue synthétique de la répartition des classes de mots-clés en fonction respectivement de leur taille et du nombre de paragraphes où elles détectent leur thème.

1. **<élection/électoral/républicain>**,
   37 mots, 7 139 paragraphes :
   { abstention, assemblée, boycottage, bulletin, candidat, candidature, circonscription, consultation, démocrate, démocrate-chrétien, désignation, électeur, élection, électoral, électorat, élu, exprimé, fraude, gouverneur, liste, législatif, mandat, popularité, proportionnel, présidence, présidentiel, référendum, républicain, score, sondage, scrutin, suffrage, sénat, sénateur, urne, votant, vote }

FIG. A.1 – *Répartition des tailles des classes de mots-clés présentées*

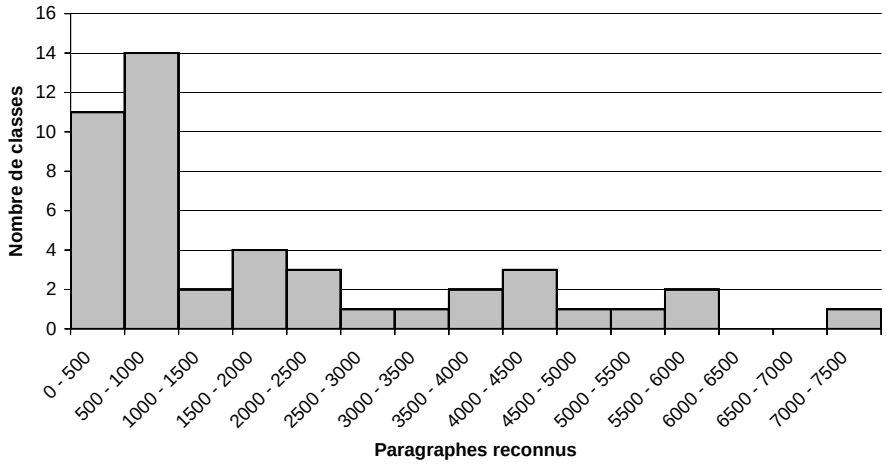


FIG. A.2 – *Répartition des nombres de paragraphes « reconnus » par les classes de mots-clés présentées*

2. **<prix/consommation/alimentaire>**,
   45 mots, 5 938 paragraphes :
   { acheteur, alimentaire, approvisionnement, autosuffisance, baril, blé, cacao, café, caoutchouc, consommateur, consommation, coton, céréale, céréalier, denrée, douane, débouché, excédentaire, exportateur, importateur, fournisseur, fruit, grain, huile, importation, lait, légume, manufacturé, matière première, maïs, pain, prix, producteur, protectionnisme, provenance, préférentiel, quota, riz, sac, stock, substitution, sucre, surplus, tabac, viande }

3. **<entreprise/secteur/investisseur>**,
   24 mots, 5 644 paragraphes :
   { actionnaire, concurrentiel, déréglementation, désengagement, entrepreneur, entreprise, immobilier, investisseur, libéralisation, lucratif, mixte, monopole, obsolète, placement, privatisation, privé, productif, rentabilité, rentable, restructuration, secteur, service public, spéculatif, spéculation }

4. **<travail/emploi/salarié>**,
   33 mots, 5 306 paragraphes :
   { confédération, congé, cotisation, embauche, emploi, employant[1], employeur, employé, flexibilité, flexible, gain, gréviste, horaire, indemnité, licenciement, mobilité, patron, patronal, patronat, payé, prime, précaire, précarité, rigidité, rémunération, salarial, salarié, syndicalisme, syndicaliste, sécurité sociale, temporaire, temps partiel, travail }

5. **<police/répression/massacre>**,
   51 mots, 4 690 paragraphes :
   { arbitraire, assassin, assassinat, assassiné, attentat, aveugle, bain, balle, barbarie, brutalité, cadavre, commando, complice, complot, déportation, émeute, enlèvement, escadron, exaction, extermination, exécuté, exécution, gendarme, génocide, horreur, impunité, manifestant, massacre, meurtre, meurtrier, mort, mystérieux, otage, police, policier, provocation, représailles, répression, sabotage, sang, sanglant, sinistre, sommaire, suspect, terreur, torture, tué, tuerie, tueur, viol, voleur }

6. **<dette/crédit/prêt>**,
   36 mots, 4 417 paragraphes :
   { annulation, bailleur, balance, bancaire, banque mondiale, banquier, caisse, comptable, créance, créancier, crédit, dette, débiteur, déficitaire, dépôt, échéance, emprunt, endettement, endetté, épargne, excédent, externe, facture, faillite, financement, liquidité, montant, net, octroi, paiement, produit national brut, prêt, remboursement, solde, trésor, versement }

[1] Occurrences du verbe *employer* reconnues par erreur comme des adjectifs au cours de l'étiquetage morphosyntaxique du corpus. D'autres erreurs similaires, que nous n'avons pas jugé bon de signaler systématiquement, sont présentes dans la suite de ces résultats.

7. **<territoire/paix/négociation>**,
   31 mots, 4 366 paragraphes :
   { annexion, autodétermination, bande, calendrier, colon, colonie, colonisation, décolonisation, entité, évacuation, hébreu, implantation, intifada, intransigeance, intégrité, intérimaire, mainmise, négociation, occupation, occupé, paix, pierre, pourparler, redéploiement, restitution, retrait, sioniste, statu quo, territoire, territorial, travailliste }

8. **<dépense/réduction/budget>**,
   30 mots, 4 169 paragraphes :
   { abaissement, ajustement, allégement, assainissement, augmentation, austérité, budget, budgétaire, contribuable, diminution, drastique, déficit, dépense, fiscal, fiscalité, harmonisation, imposition, impôt, incitation, indirect, ménage, obligatoire, prélèvement, recette, rentrée, réduction, structurel, suppression, taxe, évasion }

9. **<télévision/image/publicité>**,
   40 mots, 3 974 paragraphes :
   { antenne, audience, audiovisuel, caméra, canal, champion, chaîne, club, câble, câblé, diffusant, diffusé, exploit, feuilleton, football, image, joueur, marketing, match, médiatique, olympique, onde, petit écran, programmation, publicitaire, publicité, radio, reportage, satellite, sport, sportif, stade, station, thématique, téléspectateur, télévisé, télévision, télévisuel, vedette, émission }

10. **<film/art/scène>**,
    48 mots, 3 974 paragraphes :
    { affiche, archive, art, artiste, artistique, attraction, audace, cassette, censure, chanson, chant, chanteur, cinéaste, cinéma, cinématographique, concert, créateur, danse, disque, documentaire, esthétique, exposition, festival, fiction, film, hommage, major, musical, musicien, musique, musée, métaphore, peintre, peinture, photographie, plastique, projection, rock, réalisateur, salle, scène, spectacle, spectateur, studio, tableau, talent, théâtre, écran }

11. **<industriel/technologie/équipement>**,
    33 mots, 3 105 paragraphes :
    { accélération, acquisition, amont, aval, aéronautique, biotechnologie, chimie, combinaison, consortium, implicite, industriel, innovation, internationalisation, licence, matériau, maîtrise, militaro-industriel, moteur, mécanique, performance, performant, pesanteur, poindre, pointe, procédé, reconversion, savoir-faire, sophistiqué, technologie, technologique, tertiaire, transfert, équipement }

12. **<taux/chute/récession>**,
    35 mots, 2 586 paragraphes :
    { augmenté, baisse, billet, bourse, boursier, change, chute, conjoncturel, devise, dévaluation, fixe, fluctuation, hausse, indicateur, indice, inflation,

krach, macro-économique, marasme, optimiste, parité, produit intérieur brut, prévision, ralentissement, redressement, reprendre, récession, spéculateur, stabilisation, stagnation, taux, transaction, trimestre, variation, vertigineux }

13. **<technique/scientifique/progrès>**,
29 mots, 2 445 paragraphes :
{ académie, biologique, brevet, cerveau, chercheur, connaissance, discipline, découverte, découvrir, fonctionnel, gène, génie, génétique, ingénieur, interaction, invention, irrationnel, laboratoire, progrès, rationnel, sagesse, science, scientifique, synthèse, technicien, technique, vaccin, écologie, éthique }

14. **<guérilla/force armée/colonel>**,
34 mots, 2 081 paragraphes :
{ assaut, autodéfense, bataillon, brigade, caserne, colonel, combattant, combattre, commandant, commandement, contingent, entraîné, force armée, fusil, grenade, guérilla, guérillero, lumineux, maquis, milice, milicien, moudjahiddin, officier, opium, pacification, paramilitaire, rebelle, renfort, rébellion, sentier, subversion, trafiquant, victorieux, état-major }

15. **<nucléaire/arme/atlantique>**,
29 mots, 2 019 paragraphes :
{ agresseur, arme, atlantique, atomique, avertissement, bombe, chimique, classique, destruction, dissuasion, escalade, essai, explosif, fabrication, gel, interdiction, mutuel, nucléaire, plutonium, prolifération, riposte, réacteur, signataire, superpuissance, ultime, unilatéral, uranium, vecteur, vérification }

16. **<électronique/télécommunication/machine>**,
31 mots, 1 903 paragraphes :
{ artificiel, automatique, automobile, autoroute, composant, électronique, fabricant, fusion, informatique, intelligence, intelligent, logiciel, machine, multimédia, numérique, opérateur, ordinateur, outil, standard, support, temps réel, transmission, télécommunication, téléphone, téléphonique, téléviseur, usager, utilisateur, vidéo, virtuel, vitesse }

17. **<maladie/médical/malade>**,
36 mots, 1 692 paragraphes :
{ accident, assurance, avortement, cancer, clinique, consentement, couverture, diagnostic, docteur, décès, enfance, épidémie, espérance, expérimentation, fléau, hospitalier, hôpital, incidence, infantile, lit, malade, maladie, malnutrition, mortalité, médecine, médical, médicament, patient, pharmaceutique, profession, prévention, sanitaire, sida, test, thérapeutique, thérapie }

18. **<prison/juge/avocat>**,
28 mots, 1 679 paragraphes :
{ accusé, amnistie, aveu, avocat, cellule, code, commis, condamnation,

condamné, coupable, délinquant, délit, détention, détenu, emprisonnement, infraction, instruction, judiciaire, juge, juridiction, magistrat, plainte, prison, prisonnier, procureur, préventif, pénal, verdict }

19. **<enseignement/supérieur/scolaire>**,
35 mots, 1 616 paragraphes :
{ adulte, alphabétisation, analphabète, apprentissage, bibliothèque, collège, cycle, diplôme, diplômé, défavorisé, éducatif, élève, enseignant, enseignement, enseigné, établissement, étudiant, filière, garçon, instituteur, jeunes, lycée, lycéen, mathématique, primaire, pédagogique, qualification, recrutement, scolaire, scolarisation, secondaire, section, supérieur, sélection, titulaire }

20. **<nation/sanction/embargo>**,
21 mots, 1 376 paragraphes :
{ agression, assemblée, blocus, cessez-le-feu, chapitre, charte, communauté internationale, consultatif, égide, embargo, ingérence, intergouvernemental, levée, nation, résolution, sanction, secrétariat, session, tenue, unir, veto }

21. **<missile/spatial/arsenal>**,
28 mots, 1 264 paragraphes :
{ adverse, antimissile, arsenal, artillerie, balistique, base, bouclier, canon, char, conventionnel, croisière, déploiement, engin, équipé, flotte, fusée, hélicoptère, laser, missile, opérationnel, orbite, portée, précision, rayon, sous-marin, spatial, supériorité, tir }

22. **<forêt/sol/déchet>**,
26 mots, 987 paragraphes :
{ arbre, bois, charbon, climatique, combustible, déchet, désertification, écologique, écologiste, écosystème, électricité, énergétique, érosion, forestier, forêt, incendie, pluie, pollué, pollution, rivière, serre, sol, sûreté, toxique, tropical, végétal }

23. **<nazi/historien/fascisme>**,
50 mots, 945 paragraphes :
{ agressif, agressivité, antisémitisme, autoritaire, autoritarisme, barbare, bénin, communisme, credo, diable, dictateur, dictature, dénoncé, exaltation, extrémiste, fascisme, fasciste, fureur, haine, hindou, historien, idéologue, implacable, impérialisme, martial, maréchal, meeting, multipartisme, mystique, mythique, nationalisme, nationaliste, nazi, nazisme, négation, philosophe, philosophique, pluraliste, populisme, populiste, primitif, putsch, rassemblement, réveil, stalinien, structurer, théoricien, totalitarisme, tyrannie, uniforme }

24. **<écrivain/personnage/roman>**,
19 mots, 939 paragraphes :
{ amour, collection, écriture, écrivain, éditeur, édition, humour, littéraire,

littérature, nobel, personnage, poème, poète, poésie, revue, roman, romancier, récit, traduction }

25. **<immigrer/immigration/légal>**,
30 mots, 871 paragraphes :
{ clandestin, contrebande, dénonciation, dérapage, émigration, entrer, expulsion, fantasme, flux, hispanique, illégal, immigration, immigrer, irrégulier, légal, législation, migratoire, originaire, passeport, racisme, raciste, rechercher, regroupement, résident, souche, suspicion, sécuritaire, visa, xénophobe, xénophobie }

26. **<front/militant/libération>**,
27 mots, 850 paragraphes :
{ activiste, aile, bourgeois, djihad, exil, front, indépendantiste, insurrection, interruption, intransigeant, libération, marxisme, marxiste, militant, modéré, opposé, oppression, patriotique, progressiste, ressentiment, révolutionnaire, salut, soulèvement, subversif, tigre, trêve, unifier }

27. **<ressource/mine/pêche>**,
27 mots, 803 paragraphes :
{ abondance, bateau, conservation, créativité, cuivre, diamant, disponibilité, disponible, épuisement, extraction, gisement, mine, minerai, minier, métal, naturel, poisson, prometteur, pêche, pêcheur, renouvelable, ressource, richesse, satisfaction, solvable, suffisant, urgent }

28. **<langue/autochtone/francophone>**,
25 mots, 747 paragraphes :
{ anglophone, assimilation, autochtone, basque, coexistence, cohabitation, distinct, francophone, francophonie, fédéralisme, identitaire, langue, latin, laïcité, linguistique, maternel, mélange, néerlandais, parler, repli, slave, spécificité, unitaire, usage, vocabulaire }

29. **<liberté/respect/libre>**,
25 mots, 713 paragraphes :
{ alternance, amendement, circulation, clause, dignité, égalitaire, égalité, entrave, épanouissement, équité, expression, fondamental, fraternité, garer, idéal, individualisme, liberté, libre, marchandise, néolibéralisme, pluralisme, respect, restriction, tolérance, totalitaire }

30. **<islamique/république/fondamentaliste>**,
27 mots, 694 paragraphes :
{ accession, affaiblir, allégeance, baasiste, chiite, clergé, constitutif, coran, croire, dignitaire, divin, désintégration, fondamentaliste, fédération, imam, instauration, islamique, kurde, mosquée, propagande, prophète, ratification, république, sacré, sunnite, sécession, séparatiste }

31. **<catholique/évêque/prêtre>**,
14 mots, 666 paragraphes :
{ cardinal, catholique, christ, christianisme, évêque, laïc, missionnaire, pape, prière, prêtre, théologie, théologien, théologique, vatican }

32. **<journal/publication/lecteur>**,
26 mots, 641 paragraphes :
{ actualité, colonne, commentaire, courrier, diffusion, écrire, éditorial, éditorialiste, hebdomadaire, inédit, journal, lecteur, longueur, magazine, mensuel, organe, papier, paraître, photo, publication, périodique, rédacteur, rédaction, tirage, titre, tribune }
33. **<réfugié/asile/convention>**,
20 mots, 559 paragraphes :
{ accueil, afflux, asile, cherchant, clandestinité, commissariat, convention, demandeur, déplacé, exilé, obtention, office, persécution, rapatriement, refuge, ressortissant, restrictif, réfugié, séjour, tolérer }
34. **<réforme/soviet/foncier>**,
26 mots, 526 paragraphes :
{ agraire, baron, bourgeoisie, bureaucratique, centraliser, clef, coopératif, décret, foncier, féodal, indemnisation, initial, nationalisation, nomenklatura, oligarchie, perestroïka, planification, redistribution, réformateur, réforme, réformiste, scission, social-démocratie, soviet, terrien, timide }
35. **<animal/rendement/récolte>**,
20 mots, 515 paragraphes :
{ agriculteur, aliment, animal, bétail, bête, coopérative, élevage, engrais, fermier, intensif, irrigation, parcelle, pesticide, rendement, récolte, semence, superficie, vache, variété, vivrier }
36. **<étude/renseignement/spécialiste>**,
27 mots, 440 paragraphes :
{ amiral, anticommuniste, collaboration, colloque, confidentiel, directeur, durer, désinformation, étude, exposé, fondation, franco-allemand, institut, lever, ligue, mouvance, nommer, organisateur, participant, participant, pétition, rassemblé, renseignement, sociologique, spécialiste, séminaire, tabou }
37. **<pétrolier/royaume/saoudite>**,
16 mots, 418 paragraphes :
{ belligérant, centraméricain, différend, dynastie, émirat, frontalier, hégémonique, israélo-arabe, israélo-palestinien, litige, monarchie, pavillon, péninsule, pétrolier, royaume, saoudite }
38. **<long/contrat/chômeur>**,
35 mots, 416 paragraphes :
{ adapté, âgé, aléatoire, aménagement, bénéficiaire, charge, chômeur, collectivité, commissaire, complémentaire, contrat, court, dispositif, dividende, durée, déterminé, enthousiasme, éventail, exclusion, fixé, garanti, illimité, impopulaire, insertion, long, multiplicité, prestation, revenu, stable, stage, sélectif, terme, transitoire, utilité, volet }
39. **<quartier/architecte/habitat>**,
20 mots, 338 paragraphes :

{ agglomération, architecte, boutique, béton, centre-ville, chauffeur, dégradé, étage, fenêtre, habitat, habitation, immeuble, jardin, mairie, quartier, restaurant, rénovation, taxi, toit, villa }

40. **<apartheid/racial/sexe>**,
16 mots, 238 paragraphes :
{ abolition, apartheid, bantoustan, cap, colonialisme, discrimination, ethnique, masculin, métis, préjugé, race, racial, sexe, ségrégation, séparé, tribal }

41. **<main-d'oeuvre/vêtement/textile>**,
21 mots, 197 paragraphes :
{ abondant, acier, atelier, bon marché, chaussure, comparatif, compétitivité, dragon, dumping, délocalisation, globalisation, main-d'oeuvre, multinational, multinationale, onéreux, produire, sidérurgie, sous-traitance, textile, valeur ajoutée, vêtement }

Selon les règles de filtrage présentées à la section 2.7, les classes suivantes sont écartées du résultat final.

42. **<démographique/paysannerie/disparité>**,
22 mots, 161 paragraphes :
{ accélérer, appauvrissement, citadin, côtier, descendre, disparité, décrit, démographie, démographique, déshériter, exode, industrialisation, intensification, migrant, migration, paysannerie, peuplé, poussé, sous-développement, sous-développé, surpeuplé, urbanisation }

43. **<salon/repas/divorce>**,
31 mots, 118 paragraphes :
{ allocation, appartement, bébé, caché, chaud, coin, dame, divorce, domicile, finir, gare, gratuit, heureux, jeune femme, jeune fille, loyer, maigre, mari, mariage, natal, pension, plat, préparer, repas, retraité, salon, sans-abri, superbe, supermarché, vin, époux }

44. **<quantitatif/oeuvre/associatif>**,
29 mots, 83 paragraphes :
{ adressé, aire, analogue, antérieur, articulation, associatif, bond, cohérence, concevoir, contribuer, envisager, gravité, groupement, générateur, illicite, informer, leadership, mettre, modifier, nouveauté, oeuvre, ponctuel, prépondérant, qualitatif, quantitatif, réciproque, sectoriel, tissu, vitalité }

45. **<royal/fonction publique/affaire étrangère>**,
25 mots, 80 paragraphes :
{ affaire étrangère, ambassade, anticommunisme, attribution, cabinet, communiqué, constructif, diplomate, envoyer, fonction publique, guise, ministériel, multipliant, nettoyage, nomination, nordique, portefeuille,

représailles, restreint, royal, répété, réélection, sous-secrétaire, suspension, vice-premier }

46. **<après-guerre/rude/fief>**,
21 mots, 35 paragraphes :
{ aigu, après-guerre, artisanal, ascension, camion, dépression, feuille, fief, gamme, implosion, imprévisible, incontrôlable, irrésistible, pan, prolétariat, prospère, rationalisation, rude, rémunéré, scrupule, séance }

Annexe B

Exemples de classes sémantiques construites par analyse de l'intégralité du corpus du *Monde diplomatique*

Les quelques classes sémantiques présentées ici ont pour objectif de donner au lecteur une idée du type de résultat obtenu par la méthode de classification sémantique des mots sur la totalité du corpus présentée à la section 3.4. Elles ne constituent pas un recensement exhaustif des classes qu'il est possible d'extraire manuellement des arbres de classification produits, mais une sélection à visée purement illustrative. Pour chacune des catégories morphosyntaxiques étudiées (noms, verbes, adjectifs, adverbes et nombres), nous présentons donc les classes qui nous ont « sauté aux yeux » lors d'une lecture superficielle des arbres de classification des mots apparaissant au moins 200 fois dans le corpus intégral du *Monde diplomatique*.

Précisons que l'ordre dans lequel les mots classés sont reportés est celui dans lequel ils apparaissent dans la représentation graphique de l'arbre ; il est ainsi possible de remarquer au sein de classes les plus étendues des paires de mots successifs de sens particulièrement proches, qui correspondent le plus souvent en effet à des regroupements « de bas niveau » effectués par l'algorithme de classification.

Noms

{pour cent, nombre, millier, centaine, dizaine}

{moitié, tiers, quart}

{totalité, quasi-totalité}

{voix, vote, suffrage}

{année, période, phase, époque, ère, temps, siècle, décennie}

{an, jour, mois, août, printemps, automne, lendemain, anniversaire, veille, moment, heure, semaine}

{fin, début}

{gouvernement, président, ministre}

{population, personne, habitant, réfugié, prisonnier, détenu, enfant, femme}

{opposition, droite, gauche}

{citoyen, gens, Américain, Français, Israélien, Russe, Britannique, Albanais, Palestinien, Indien, Arabe, juif, Allemand, Africain, musulman, Serbe}

{paysan, travailleur, salarié, ouvrier, agriculteur, entrepreneur}

{organisation, association, syndicat}

{système, structure, mécanisme, dispositif, méthode, technique, instrument, outil}

{politique, stratégie}

{projet, programme, plan}

{entreprise, banque, firme, compagnie, géant, concurrent, multinationale}

{secteur, domaine}

{démocratie, capitalisme, libéralisme}

{région, zone, sud, nord, est, ouest, tiers-monde, golfe, mer, continent, proche-orient}

{ville, province, voisin, île, cité, capitale, métropole, agglomération}

{hectare, surface}

{nuit, journée, soir, matin}

{village, quartier, banlieue, maison, hôtel, appartement}

{armée, troupe, soldat, militaire, officier, police, corps, garde}

{guerre, conflit, affrontement, confrontation, rivalité, querelle}

{lutte, combat}

{résistance, guérilla, rébellion, révolte, insurrection, soulèvement, émeute, trouble}

{manifestation, grève}

{crise, événement, scandale, accident, incident, catastrophe, désastre, drame, tragédie}

{violence, corruption, racisme, délinquance, criminalité}

{tension, inquiétude, mécontentement, désarroi, malaise, incertitude, insécurité}

{récession, sécheresse, famine, chaos, désordre}

{différence, inégalité, disparité}

{convergence, divergence}

{acte, crime, délit, assassinat, meurtre, attentat, massacre}

{sport, football}

{père, mère, mari, époux, parent, frère, fils}

{classe, couche, caste, élite, bourgeoisie}
{témoignage, révélation, photo}
{impression, indifférence, optimisme}
{retard, avance}
{délai, échéance}
{reprise, arrêt, poursuite}
{sanction, embargo, blocus, restriction, limitation, interdiction}
{scène, théâtre, spectacle}
{appareil, machine, ordinateur}
{essence, alcool}
{échelle, échelon, rang}
{client, fournisseur, producteur, consommateur, utilisateur}
{concurrence, bataille, compétition}
{répression, dissuasion, riposte, représailles}
{attaque, agression, occupation, invasion}
{gestion, régulation, planification}
{chute, effondrement, disparition, explosion, déclin, essor, destruction, amélioration, dégradation, détérioration, renforcement, affaiblissement}
{faiblesse, insuffisance, précarité, fragilité}
{ouverture, privatisation, libéralisation, déréglementation}
{transition, démocratisation, normalisation}
{appel, recours, usage, utilisation}
{maintien, instauration, rétablissement, restauration}
{illusion, façade}
{membre, chef, secrétaire, vice-président, directeur, directeur général, maire, gouverneur, sénateur, tête, représentant}
{professeur, enseignant, étudiant, élève, ingénieur}
{journaliste, artiste, écrivain, personnalité, intellectuel}
{expert, observateur, économiste, historien, chercheur, spécialiste, médecin, avocat, juriste}
{conseil, commission, comité, congrès, parlement, chambre, sénat}
{institution, organisme, instance, organe}
{texte, document, déclaration, résolution, lettre, message}
{idéologie, doctrine, dogme, thèse, théorie}
{efficacité, audience, rentabilité, productivité, compétitivité, performance}
{céréale, blé, riz, maïs, coton, sucre, viande}
{situation, environnement, contexte, climat, atmosphère, circonstance, conjoncture}
{négociation, discussion, pourparler, conversation}
{phrase, chapitre}

Verbes

{falloir, vouloir}

{exprimer, manifester, confirmer, témoigner}

{provoquer, entraîner, accompagner, accélérer, freiner, enrayer, aggraver, accentuer, engendrer, éviter, empêcher}

{atteindre, augmenter, diminuer, accroître, élever, baisser, chuter, progresser, dépasser, monter, alimenter}

{favoriser, nécessiter, faciliter}

{impliquer, signifier, supposer}

{exiger, réclamer, préconiser, prôner}

{expliquer, souligner, noter, remarquer, constater, observer}

{mesurer, évaluer}

{prouver, démontrer}

{refléter, illustrer}

{désigner, nommer}

{mourir, tuer, assassiner}

{construire, détruire}

{employer, recruter, licencier}

{perdre, garder, conserver}

{encourager, relancer, stimuler}

{réaliser, effectuer, opérer}

{remplacer, substituer}

{aboutir, déboucher}

{vendre, acheter}

{coûter, rapporter}

{causer, compromettre, conditionner}

{entamer, amorcer}

{défendre, combattre, lutter}

{craindre, redouter}

{essayer, efforcer, songer}

{étonner, surprendre}

{maintenir, préserver, rétablir, consolider}

{rompre, briser}

{proposer, adopter, élaborer}

{signer, conclure, proclamer, adhérer, ratifier}

{raconter, décrire, étudier, examiner, analyser, résumer, exposer}

Adjectifs

{politique, économique, social}

{américain, français}

{premier, dernier, prochain, précédent, récent, second, troisième, deuxième}

{bon, mauvais}

{ancien, vieux}

{noir, blanc, rouge, vert}

{petit, gros}

{principal, essentiel, fondamental, majeur}

{international, mondial, régional}

{important, considérable, énorme, immense, gigantesque}

{net, significatif, substantiel, notable}

{formidable, extraordinaire, remarquable, étonnant}

{total, complet}

{massif, systématique, continu, constant, rapide, progressif, brutal}

{démocratique, populaire, civique, patriotique, militant}

{vrai, faux}

{technique, technologique}

{intellectuel, artistique, littéraire}

{collectif, personnel, individuel}

{financier, monétaire, budgétaire, fiscal, foncier}

{nucléaire, atomique}

{monter, naval, automobile, chimique, textile, agroalimentaire}

{électronique, électrique, téléphonique, audiovisuel, cinématographique}

{structurel, institutionnel}

{judiciaire, policier, pénal}

{européen, étranger, occidental, soviétique, israélien, irakien, syrien, saoudien, pakistanais, iranien, turc, indonésien, argentin, marocain, sud-africain, cubain, vietnamien, allemand, russe, japonais, britannique, italien, espagnol, anglais, portugais, palestinien, serbe, croate, éthiopien, africain, latino-américain, brésilien, mexicain, belge, yougoslave, bosniaque, canadien, algérien, libanais, afghan, grec, égyptien, polonais, hongrois, chinois, indien}

{communiste, nazi, fasciste, nationaliste, islamiste, indépendantiste, écologiste, islamique, marxiste, laïque, libéral, néolibéral, radical, révolutionnaire, favorable, hostile, conservateur, extrémiste, fondamentaliste, modérer, opposer, anticommuniste, socialiste, démocrate, républicain, travailliste, social-démocrate, sortant}

{juif, catholique, orthodoxe, musulman, chrétien, kurde, chiite, indigène, hindou}

{religieux, ethnique, racial, linguistique, territorial}

{industrialisé, développé, prospère, avancé}
{multinational, transnational}
{dominant, majoritaire, minoritaire}
{droit, gauche}
{long, bref, court}
{cinquième, quatrième, dix-neuvième, vingtième}
{fameux, célèbre}
{moyen, supérieur, inférieur, haut, bas, faible, suffisant, insuffisant}
{égal, équivalent, normal, ordinaire}
{annuel, mensuel, hebdomadaire, régulier}
{irréversible, inéluctable}
{grave, terrible, dramatique, tragique, désastreux, catastrophique}
{définitif, provisoire, temporaire}
{intéressant, frappant, révélateur, marquant, caractéristique}
{autorisé, interdit}

Adverbes

{particulièrement, d'autant plus}
{souvent, parfois}
{aujourd'hui, désormais}
{immédiatement, aussitôt}
{rapidement, progressivement}
{totalement, complètement}
{régulièrement, fréquemment, sans cesse, constamment}
{surtout, notamment}
{essentiellement, principalement, uniquement, exclusivement}
{pourtant, cependant, toutefois}
{officiellement, ouvertement, publiquement}
{sans doute, peut-être, évidemment, probablement, nécessairement}
{difficilement, aisément}
{parallèlement, simultanément}
{près, environ, quelque, presque, quasi, quasiment, pratiquement}
{vraiment, réellement, sérieusement}
{à peine, rarement}
{fortement, profondément, considérablement, sensiblement}
{précisément, juste, justement}
{tard, tôt}
{autrefois, jadis, naguère}
{nettement, radicalement, fondamentalement}

{durement, brutalement}
{strictement, étroitement}

Nombres

{deux, trois, Deux, Trois, quatre, cinq}
{300 000, 500 000, 100 000, 200 000}
{1996, 1997, 1945}
{1990, 1992, 1994, 1995, 1991, 1993, 1989, 1988, 1984, 1987, 1985, 1986, 1982, 1981, 1979, 1974, 1975, 1976, 1973, 1970, 1978, 1977, 1980, 1983}
{70, 75, 55, 65, 60, 50, 40, 45, 35}
{20, 22, 21, 18, 19, 27, 24, 16, 23, 28, 26, 11, 17, 9, 14, 13, 15, 30}
{vingt, quarante, vingt-cinq, cinquante, treize, trente, dix-huit, vingt-quatre, seize, soixante}
{2.5, 1.5}

Annexe C

Glossaire

Sont rassemblées ici les définitions des termes spécialisés employés (parfois « en passant ») dans cette thèse, qu'il s'agisse de termes linguistiques (les définitions doivent dans ce cas beaucoup au glossaire publié par F. Rastier dans [Ras89] ainsi qu'à celui de P. Vaillant dans [Vai99]), statistiques ou purement techniques. Nous y faisons également figurer les dénominations ou notations que nous avons introduites dans le cadre de notre travail.

acception : sémème* dont le sens comprend des sèmes* afférents socialement normés.

actualisation : opération interprétative permettant d'identifier ou de construire un sème* en contexte.

afférence : inférence permettant d'actualiser un sème* afférent.

champ : ensemble des taxèmes* pertinents dans une tâche.

classe de mots-clés : dans le cadre de notre travail (chapitre 2), ensemble de mots-clés* révélateurs d'un même thème, permettant à la fois de le caractériser et d'en détecter les occurrences.

classe sémantique : une préoccupation courante de l'acquisition automatique d'informations lexicales est la construction de classes sémantiques rassemblant des mots de sens proches (typiquement, pouvant être substitués les uns aux autres dans certains contextes). Dans le cadre de nos travaux (chapitre 3), les classes sémantiques construites rejoignent la notion théorique de taxème*.

classème : ensemble des sèmes* génériques d'un sémème*.

clique : partie d'un graphe tel qu'il existe un arc entre chaque nœud et tous les autres.

complexe sémique : structure sémantique temporaire qui résulte de l'assemblage des sémies* dans le syntagme (par activation* et virtualisation* de sèmes*, *etc.*). Au palier textuel, les complexes sémiques analogues sont considérés comme des occurrences de la même molécule* sémique.

composant : trait sémantique. On distingue deux sortes de composants, les sèmes* et les primitives*.

composante : instance systématique qui participe à régler la production et l'interprétation des suites linguistiques. Pour le plan du contenu, on distingue quatre composantes : thématique*, dialectique*, dialogique* et tactique*.

contexte : pour une unité sémantique, ensemble des unités qui ont une incidence sur elle (contexte actif), et sur lequel elle a une incidence (contexte passif). Le contexte connaît autant de zones de localité qu'il y a de paliers* de complexité.

dialectique : composante sémantique qui articule la succession des intervalles dans le temps textuel, comme les états qui y prennent place et les processus qui s'y déroulent.

dialogique : composante sémantique qui articule les relations modales entre univers et entre mondes.

dimension : classe de sémèmes de généralité supérieure, indépendante des domaines. Les dimensions sont groupées en petites catégories fermées (ex : // animé // *vs* // inanimé //).

discours : ensemble d'usages linguistiques codifiés attaché à un type de pratique sociale. Ex. : discours juridique, médical, religieux.

domaine : groupe de taxèmes lié à une pratique sociale. Il est commun aux divers genres propres au discours qui correspond à cette pratique. Dans un domaine déterminé, il n'existe généralement pas de polysémie.

entropie : mesure de la complexité ou « imprédictabilité » d'un signal, l'une des bases de la théorie de l'information [Sha48]. L'entropie d'une variable aléatoire X est définie par :

$$H(X) = -\sum_{x} P(x).\log_2\left(P(x)\right)$$

où $P(x)$ est la probabilité pour la variable X d'être dans l'état x, et en posant $p.\log_2(p) = 0$ si $p = 0$.

information mutuelle : l'information mutuelle [CH89] est fréquemment utilisée en traitement automatique des langues afin de mesurer l'« attirance » pouvant exister entre deux mots dans un texte, révélant potentiellement l'existence d'un lien syntagmatique* entre eux. Elle se définit dans ce contexte par :

$$I\left(m_1, m_2\right) = \log_2 \frac{P\left(m_1, m_2\right)}{P\left(m_1\right).P\left(m_2\right)}$$

où $P\left(m_1, m_2\right)$ est défini comme « la probabilité de voir apparaître simultanément m_1 et m_2 dans un même contexte », la définition du mot « contexte » étant amenée à varier.

interprétation : assignation d'un sens à une suite linguistique.

isotopie sémantique : effet de la récurrence d'un même sème ou groupe de sèmes. L'isotopie est selon F. Rastier à l'origine de l'impression de cohérence d'un texte, et constitue le principal guide du processus interprétatif.

Jaccard (indice de) : indice numérique simple mesurant la similarité entre deux ensembles d'objets. A et B étant deux ensembles quelconques, l'indice de Jaccard entre A et B est défini par :

$$J(A,B) = \frac{\mathrm{Card}\,(A \cap B)}{\mathrm{Card}\,(A \cup B)}$$

et se situe toujours dans l'intervalle $[0,1]$. En analyse statistique de données, A et B sont typiquement des ensembles d'attributs caractérisant deux objets à comparer.

Kullback-Leibler (divergence de) : la divergence de Kullback-Leibler [KL51] entre deux distributions de probabilités p et q est définie par (dans le cas discret) :

$$\sum_x p(x)\ \log\frac{p(x)}{q(x)}$$

et est égale à la différence entre l'entropie* croisée de p et q et l'entropie de p. On rencontre également l'appellation d'« entropie relative de p et q », ou de « distance de Kullback-Leibler », bien que la mesure définie ne soit pas symétrique et ne respecte pas l'inégalité triangulaire.

lexicographie : partie de la linguistique appliquée qui se consacre à la rédaction de lexiques ou dictionnaires.

lexicologie : étude linguistique du lexique.

lexie : groupement stable de morphèmes, constituant une unité fonctionnelle.

Manhattan (métrique de) : la métrique de Manhattan correspond à la notion topologique de « norme infinie », et définit la distance entre deux vecteurs comme la somme des valeurs absolues des différences de leurs coordonnées. Elle est ainsi nommée car elle peut s'interpréter comme la distance d'un point à un autre dans une ville où toutes les rues se coupent à angle droit.

molécule sémique : groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier. Un « thème », quand il peut être défini sémantiquement, n'est autre qu'une molécule sémique.

morphème : Plus petite partie indécomposable d'un mot dotée d'un sens. Par exemple, le mot *rétropropulseurs* compte cinq morphèmes : *rétro*, *pro*, *puls*, *eur*, et *s*.

mot : groupement de morphèmes complètement intégré.

mot-clé : dans le cadre de nos travaux (chapitre 2), mot fortement révélateur d'un thème donné apparaissant fréquemment dans le discours développé dans le corpus autour de ce thème. Regroupés en classes de mots-clés*, ils permettent la caractérisation et la détection des occurrences des thèmes du corpus.

ordre syntagmatique : ordre de la linéarisation du langage, dans une étendue spatiale ou temporelle. Il rend compte des relations positionnelles et des relations fonctionnelles. Ainsi, il est le site des relations contextuelles.

ordre paradigmatique : ordre de l'association codifiée. Une unité sémantique ne prend sa valeur que relativement à d'autres qui sont substituables avec elle et qui forment son paradigme de définition.

ordre herméneutique : ordre des conditions de production et d'interprétation des textes. Il englobe les phénomènes de communication, mais dépasse les facteurs pragmatiques, en incluant les situations de communication codifiées, différées, et non nécessairement interpersonnelles. Il est inséparable des situations historique et culturelle de la production et de l'interprétation.

ordre référentiel : ordre qui détermine l'incidence du linguistique sur les strates non linguistiques de la pratique. Il participe à la constitution d'impressions référentielles.

palier : degré de complexité, ou échelle textuelle, pouvant aller du morphème au texte dans sa totalité.

parcours interprétatif : suite d'opérations permettant d'assigner un ou plusieurs sens à une suite linguistique.

référence : être ou état de choses du « monde réel » que veut évoquer un mot ou un syntagme. La sémantique et la logique modernes se sont largement développées sur la distinction établie par G. Frege dans [Fre92] entre sens* et référence.

réseau associatif : ensemble des relations qui permettent d'identifier la récurrence d'une molécule sémique.

sémantème : ensemble des sèmes spécifiques d'un sémème.

sémème : signifié d'un morphème.

sème : élément d'un sémème, défini comme l'extrémité d'une relation fonctionnelle binaire entre sémèmes. Le sème est la plus petite unité de signification définie par l'analyse.

sème afférent : sème associé à un sémème par une norme sociale, et non par le système de la langue, par exemple / faiblesse / pour *femme*. Contrairement au sème inhérent*, un sème afférent n'est en théorie pas présent dans le lexique, mais est actualisé par instruction contextuelle.

sème générique : élément du classème, marquant l'appartenance du sémème à une classe sémantique (taxème, domaine, ou dimension).

sème inhérent : sème que l'occurrence hérite du type, par défaut. Un exemple courant est / noir / pour *corbeau*.

sème spécifique : élément du sémantème* opposant le sémème à un ou plusieurs sémèmes du taxème auquel il appartient. Par exemple, / sexe féminin / distingue *femme* de *homme*.

sémie : signifié d'une lexie.

sens : le sens doit être distingué, d'une part, de la référence*, car il est défini au sein du système de la langue, sans référence au « monde réel », et d'autre part de la signification*, qui est définie pour un mot de manière statique et absolue tandis que son sens ne prend corps qu'en contexte, par l'usage. Selon la sémantique interprétative, le sens d'un mot est défini par l'ensemble des sèmes actualisés pour ce mot dans un contexte d'apparition donné.

signification : signifié d'une unité linguistique, défini en faisant abstraction des contextes et des situations. Toute signification est ainsi un artefact.

signifié : contenu d'une unité linguistique.

taxème : ensemble restreint de signes s'interdéfinissant par des oppositions minimales. Le taxème reste en général de l'ordre de l'immanence, chaque signe présent dans un texte ayant justement été choisi de préférence à tous les autres membres du même taxème — qui auraient pu éventuellement se trouver à la même place.

texte : suite linguistique autonome (orale ou écrite) constituant une unité empirique, et produite par un ou plusieurs énonciateurs dans pratique sociale attestée. « Les textes sont l'objet empirique de la linguistique. Isoler des phrases, *a fortiori* des mots, résulte d'une décision méthodologique » [Ras96].

ultramétrique : métrique d vérifiant une forme renforcée de l'inégalité triangulaire, soit $\forall(a,b,c),\ d(a,c) \leq \max(d(a,b), d(b,c))$. En outre, deux des trois distances $d(a,b)$, $d(a,c)$ et $d(b,c)$ sont égales.

virtualisation : opération interprétative consistant à « étouffer » en contexte un sème inhérent au sens d'un mot dans le lexique.

Bibliographie

[ABR95] Susan Armstrong, Pierrette Bouillon, et Gilbert Robert. Tagger Overview. Rapport de recherche, ISSCO, Genève, Suisse, 1995.

[Aga95] Rajeev Agarwal. *Semantic Feature Extraction from Technical Texts with Limited Human Intervention*. Phd thesis in computer science, Mississipi State University, EU, 1995.

[AN83] Antoine Arnault et Pierre Nicole. *La logique ou l'art de penser*. Guillaume Desprez, Paris, France, 5e édition, 1683.

[Arm96] Susan Armstrong. Multext: Multilingual Text Tools and Corpora. In *Lexikon und Text*, Niemeyer, Allemagne, 1996.

[Aud03] Laurent Audibert. *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. Thèse de doctorat en informatique, Université de Provence – Aix-Marseille I, Marseille, France, 2003.

[Bar73] Roland Barthes. Texte (théorie du). In *Encyclopedia Universalis, tome XV*. Encyclopedia Universalis, Paris, France, 1973.

[BdMEBS97] Brigitte Bigi, Renato de Mori, Marc El-Bèze, et Thierry Spriet. Combined Models for Topic Spotting and Topic-Dependent Language Modeling. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 97)*, Santa Barbara, CA, EU, 1997.

[Beu98] Pierre Beust. *Contribution a un modele interactionniste du sens – Amorce d'une compétence interprétative pour les machines*. Thèse de doctorat en informatique, Université de Caen, Caen, France, 1998.

[BHDM01] Slim Ben-Hazez, Jean-Pierre Desclés, et Jean-Luc Minel. Modèle d'exploration contextuelle pour l'analyse sémantique de textes. In *8e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, Tours, France, 2001.

[BHNZ97] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, et Pierre Zweigenbaum. Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies. In *Premières JST FRANCIL 97*, Avignon, France, 1997.

[BHNZ00] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, et Pierre Zweigenbaum. Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. In Jean Charlet, Manuel Zacklad, Gilles Kassel, et Didier Bourigault, éditeurs, *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, pages 275–290. Eyrolles, Paris, France, 2000.

[BHNZ01] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, et Pierre Zweigenbaum. Corpus-Based Extension of a Terminological Semantic Lexicon. In Didier Bourigault, Christian Jacquemin, et Marie-Claude L'Homme, éditeurs, *Recent Advances in Computational Terminology*, pages 327–351. John Benjamins, Amsterdam, Pays-Bas, 2001.

[BLB01] Narjes Boufaden, Guy Lapalme, et Yoshua Bengio. Topic Segmentation: A First Stage to Dialog-Based Information Extraction. In *6th Natural Language Processing Pacific Rim Symposium (NLPRS 01)*, Tokyo, Japon, 2001.

[BN00] Branimir K. Boguraev et Mary S. Neff. Lexical Cohesion, Discourse Segmentation and Document Summarization. In *6e conférence internationale sur la Recherche d'Informations Assistée par Ordinateur (RIAO 2000)*, Paris, France, 2000.

[Bou94] Didier Bourigault. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes.* Thèse de doctorat en informatique linguistique, École des Hautes Études en Sciences Sociales, Paris, France, 1994.

[BPd+92] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, et Robert L. Mercer. Class-Based N-Gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, 1992.

[Bré97] Michel Bréal. *Essai de sémantique (science des significations)*. Hachette et Cie, Paris, France, 1897.

[Bri92] Eric Brill. A Simple Rule-Based Part of Speech Tagger. In *3rd Conference on Applied Natural Language Processing (ANLP 92)*, Trente, Italie, 1992.

[Bru03] Armelle Brun. *Détection de thèmes et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. Thèse de doctorat en informatique, Université Henri Poincaré, Nancy, France, 2003.

[Car90] Sandra Carberry. *Plan Recognition in Natural Language Dialogue*. MIT Press, Cambridge, MA, EU, 1990.

[CC99] Sharon A. Caraballo et Eugene Charniak. Determining the Specifity of Nouns from Text. In *Joint SIGDAT Conference on Empirical*

Methods in Natural Language Processing (EMNLP) and Very Large Corpora (VLC), University of Maryland, College Park, MD, EU, 1999.

[CH89] Kenneth W. Church et Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *27th Annual Meeting of the Association for Computational Linguistics (ACL 89)*, Vancouver, Canada, 1989.

[Cho57] Noam Chomsky. *Syntactic Structures*. MIT Press, Cambridge, MA, EU, 1957.

[CL04] Vincent Claveau et Marie-Claude L'Homme. Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus. In *3rd Workshop on Computational Terminology (CompuTerm'04)*, Genève, Suisse, 2004.

[Cla03] Vincent Claveau. *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. Thèse de doctorat en informatique, Université de Rennes I, Rennes, France, 2003.

[Con02] Joseph Conrad. *Heart of Darkness*. J.M. Dent & Sons, Ltd, Londres, RU, 1902.

[Cos68] Eugenio Coseriu. Les structures lexématiques. *Zeitschrift für französische Sprache und Literatur*, 1, 1968.

[Cou05] Mathieu Coutaud. Structuration automatique de classes sémantiques. Rapport de stage, master 2 recherche en informatique, IRISA, Université de Rennes I, Rennes, France, 2005.

[CR98] Eugene Charniak et Brian Roark. Noun Phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction. In *17th International Conference on Computational Linguistics (COLING 98)*, Montréal, Canada, 1998.

[Dai94] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse de doctorat en informatique, Université de Paris VII, Paris, France, 1994.

[Den04] Yves-Pol Denielou. Acquisition automatique de relations lexicales sémantiques sur corpus. Rapport de stage, ENS / IRISA, Rennes, France, 2004.

[DJ95] Bonnie J. Dorr et Doug Jones. Automatic Extraction of Semantic Classes from Syntactic Information in Online Resources. Technical report, University of Maryland, College Park, MD, EU, 1995.

[DJ96] Bonnie J. Dorr et Doug Jones. Use of Syntactic and Semantic Filters for Lexical Acquisition: Using WordNet to Increase Precision. In *Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases*, Toulouse, France, 1996.

[dS16] Ferdinand de Saussure. *Cours de Linguistique Générale*. Payot, Lausanne, Suisse, 5e édition, 1916.

[DT00] Dan Dascălu et Dan Tufiş, éditeurs. *Romanian Journal of Information Science and Technology, special edition on BalkaNet*, volume 3. Publishing House of the Romanian Academy, Bucarest, Roumanie, 2000.

[DW02] Beate Dorow et Dominic Widdows. A Graph Model for Unsupervised Lexical Acquisition. In *19th International Conference on Computational Linguistics (COLING 02)*, Taipei, Taiwan, 2002.

[ET91] Brad Efron et Robert Tibshirani. Statistical Analysis in the Computer Age. *Science*, 253:390–395, 1991.

[FBY92] William B. Frakes et Ricardo Baeza-Yates, éditeurs. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, NJ, EU, 1992.

[Fel98] Christiane Fellbaum, éditeur. *WordNet, an Electronic Lexical Database*. MIT Press, Cambridge, MA, EU, 1998.

[FG01] Olivier Ferret et Brigitte Grau. Utiliser des corpus pour amorcer une analyse thématique. *Traitement Automatique des Langues (TAL), numéro spécial Traitement Automatique des Langues et Linguistique de Corpus*, 42(2):517–545, 2001.

[FG02] Olivier Ferret et Brigitte Grau. A Bootstrapping Approach for Robust Topic Analysis. *Natural Language Enginering (NLE), Special Issue on Robust Methods of Corpus Analysis*, 8(3):209–233, 2002.

[FGMP01] Olivier Ferret, Brigitte Grau, Jean-Luc Minel, et Sylvie Porhiel. Repérage de structures thématiques dans des textes. In *8e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, Tours, France, 2001.

[FHL97] Cécile Fabre, Benoît Habert, et Dominique Labbé. La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, 13:15–30, 1997.

[FK67] Nelson Francis et Henry Kucera. *Computational Analysis of Present-day American-English*. Brown University Press, Providence, RI, EU, 1967.

[FN98] David Faure et Claire Nédellec. A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In *LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, Grenade, Espagne, 1998.

[Fol02] Helka Folch. *Articulier les classifications sémantiques induites d'un domaine*. Thèse de doctorat en informatique, Université de Paris XIII, Paris, France, 2002.

[Fre92] Gottlob Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 1892. Traduction française parue dans G. Frege, *Écrits logiques et philosophiques*, Seuil, Paris, France, 1971.

[Gre66] Algirdas Julien Greimas. *Sémantique Structurale*. Larousse, Paris, France, 1966.

[Gre93] Gregory Grefenstette. Automatic Thesaurus Generation from Raw Text Using Knowledge-Poor Techniques. In *Making Sense of Words, 9th Annual Conference of the UW Centre for the New* Oxford English Dictionary *and Text Research*, Oxford, RU, 1993.

[Gre94] Gregory Grefenstette. Corpus-derived First, Second and Third Order Word Affinities. In *6th Congress of the European Association for Lexicography (Euralex 94)*, Amsterdam, Pays-Bas, 1994.

[HA94] John Hughes et Eric Atwell. A Methodical Approach to Word Class Formation Using Automatic Evaluation. In *Workshop on Computational Linguistics for Speech and Handwriting Recognition, Convention on Artificial Intelligence and the Simulation of Behaviour (AISB 94)*, Leeds, RU, 1994.

[Har51] Zellig Harris. *Methods in Structural Linguistics*. University of Chicago Press, Chicago, IL, EU, 1951.

[Har68] Zellig Harris. *Mathematical Structures of Language*. John Wiley & Sons, New York, NJ, EU, 1968.

[HDZS01] Xiaofeng He, Chris H.Q. Ding, Hongyuan Zha, et Horst D. Simon. Automatic Topic Identification Using Webpage Clustering. In *IEEE International Conference on Data Mining*, San Jose, CA, EU, 2001.

[Hea92] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics (COLING 92)*, Nantes, France, 1992.

[Hea94] Marti A. Hearst. Multi-Paragraph Segmentation of Expository Texts. In *32th Annual Meeting of the Association for Computational Linguistics (ACL 94)*, Las Cruces, NM, EU, 1994.

[HGR+89] Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, Tzvee N. Harris, et Suzanna Harris. *The Form of Information in Science: Analysis of an Immunology Sublanguage*. Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1989.

[Hin90] Donald Hindle. Noun Classification from Predicate-Argument Structures. In *28st Annual Meeting of the Association for Computational Linguistics (ACL 90)*, Pittsburgh, PA, EU, 1990.

[HNN96] Benoît Habert, Élie Naulleau, et Adeline Nazarenko. Symbolic Word Clustering for Medium-Size Corpora. In *16th International Conference on Computational Linguistics (COLING 96)*, Copenhague, Danemark, 1996.

[Hut88] Alan Hutchinson. Building Grammars from Natural Text. In *European Working Session on Machine Learning*, Glasgow, RU, 1988.

[ICG93] *Proceedings of the 1st International Colloquium on Grammatical Inference (ICGI'93)*, Colchester, RU, 1993.

[JM04] Christian Jacquemin et Emmanuel Morin. Automatic Acquisition and Expansion of Hypernym Links. *Computer and the Humanities*, 38(4):363–396, 2004.

[JNXH05] Liping Jing, Michael K. Ng, Jun Xu, et Joshua Zhexue Huang. Subspace Clustering of Text Documents with Feature-Weighting K-Means Algorithm. In *9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 05)*, Hà Nội, Vietnam, 2005.

[KdM90] Roland Kuhn et Renato de Mori. A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, 1990.

[KGW01] Dimitrios Kokkinakis, Maria Toporowska Gronostaj, et Karin Warmenius. Corpus-Based Extension of Semantic Lexicons in Large Scale. In *13th Nordic Computational Linguistics Conference (NODALIDA 2001)*, Uppsala, Suède, 2001.

[Kil97] Adam Kilgarriff. "I don't Believe in Word Senses". *Computer and the Humanities*, 31(2):91–113, 1997.

[Kil01] Adam Kilgarrif. Generative Lexicon Meets Corpus Data: The Case of Nonstandard Word Uses. In Pierrette Bouillon et Frederica Busa, éditeurs, *The Language of Word Meaning*, pages 312–328. Cambridge University Press, Cambridge, RU, 2001.

[KL51] Solomon Kullback et Richard A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[Lau94] Andy Lauriston. Automatic Recognition of Complex terms: Problems and the TERMINO Solution. *Terminology*, 1(1):147–170, 1994.

[Ler70] Israël-César Lerman. Sur l'analyse des données préalable à une classification automatique, proposition d'une nouvelle mesure de similarité. Rapport, MSH Paris, 1970.

[Ler91] Israël-César Lerman. Foundations in the Likelihood Linkage Analysis Classification Method. *Applied Stochastic Models and Data Analysis*, 7:69–76, 1991.

[Lev93] Beth Levin. *English Verb Class and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, EU, 1993.

[LGB94] Geoffrey N. Leech, Roger Garside, et Michael Bryant. CLAWS4: the Tagging of the British National Corpus. In *15th International Conference on Computational Linguistics (COLING 94)*, Kyoto, Japon, 1994.

[LHKK96] Krista Lagus, Timo Honkela, Samuel Kaski, et Teuvo Kohonen. Self-organizing Maps of Document Collections: A New Approach to Interactive Exploration. In *Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA, EU, 1996.

[Lin98] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *17th International Conference on Computational Linguistics (COLING 98)*, Montréal, Canada, 1998.

[Lit85] Diane Litman. *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues*. Phd thesis in computer science, Rochester University, EU, 1985.

[LLP94] Israël-César Lerman, Henri Leredde, et Philippe Peter. Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien) – deuxième partie. *Revue de* MODULAD, 13:63–90, 1994.

[LP95] Diane J. Litman et Rebecca J. Passonneau. Combining Multiple Knowledge Sources for Discourse Segmentation. In *33th Annual Meeting of the Association for Computational Linguistics (ACL 95)*, Montréal, Canada, 1995.

[LP01] Dekang Lin et Patrick Pantel. Induction of Semantic Classes from Natural Language Text. In *7th International Conference on Knowledge Discovery and Data Mining (SIGKDD 01)*, San Francisco, CA, EU, 2001.

[LRDR93] Michel Legrain, Josette Rey-Debove, et Alain Rey, éditeurs. *Le Nouveau Petit Robert*. Dictionnaires Le Robert, Paris, France, 1993.

[LRS04] Mathieu Lafourcade, Frédéric Rodrigo, et Didier Schwab. Low Cost Automatic Conceptual Vector Generation from Mono and Bilingual Resources. In *Colloque de PAPILLON 2004*, Grenoble, France, 2004.

[LYG02] Winston Lin, Roman Yangarger, et Ralph Grishman. Unsupervised Learning of Generalized Names. In *19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 2002.

[LYG03] Winston Lin, Roman Yangarger, et Ralph Grishman. Bootstrapped Learning of Semantic Classes from Positive and Negative Examples. In *Workshop on The Continuum from Labeled to Unlabeled Data, 20th International Conference on Machine Learning (ICML 2003)*, Washington D.C., EU, 2003.

[MAE86] Judith Markowitz Markowitz, Thomas Ahlswede, et Martha Evens. Semantically Significant Patterns in Dictionary Definitions. In *24th Annual Meeting of the Association for Computational Linguistics (ACL 86)*, New York, NJ, EU, 1986.

[Mar67] André Martinet. *Éléments de Linguistique Générale*. Armand Colin, Paris, France, 2e édition, 1967.

[Mar80] Mitchell P. Marcus. *A Theory of Syntactic Recognition for Natural Language*. The MIT Press, Cambridge, MA, EU, 1980.

[MC91] George A. Miller et William G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[MCP95] Igor Mel'cuk, André Clas, et Alain Polguère. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve, Belgique, 1995.

[MS96] John G. McMahon et Francis J. Smith. Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies. *Computational Linguistics*, 22(2):217–247, 1996.

[MS01] Paola Merlo et Suzanne Stevenson. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408, 2001.

[MS05] Fabienne Moreau et Pascale Sébillot. Contribution des techniques du traitement automatique des langues à la recherche d'informations. Publication interne 1690, IRISA, Rennes, France, 2005.

[MZB04] Véronique Malaisé, Pierre Zweigenbaum, et Bruno Bachimont. Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologies. In *11e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 04)*, Fès, Maroc, 2004.

[NN88] Jun-ichi Nakamura et Makoto Nagao. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. In *12th International Conference on Computational Linguistics (COLING 88)*, Budapest, Hongrie, 1988.

[NV02] Roberto Navigli et Paola Velardi. Automatic Adaptation of WordNet to Domains. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Îles Canaries, Espagne, 2002.

[NV04] Roberto Navigli et Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151–179, 2004.

[OM97] Akira Oishi et Yuji Matsumoto. Detecting the Organization of Semantic Subclasses of Japanese Verbs. *International Journal of Corpus Linguistics*, 1(2):65–89, 1997.

[Per04] Vincent Perlerin. *Sémantique légère pour le document*. Thèse de doctorat en informatique, Université de Caen, Caen, France, 2004.

[Pla67] Platon. Protagoras. In *Protagoras – Euthydème – Gorgias – Ménexène – Ménon – Cratyle*. GF-Flammarion, Paris, France, 1967. trad. Émile Chambry.

[Pot62] Bernard Pottier. *Systématique des éléments de relation. Étude de morphosyntaxe structurale romane*. Klincksieck, Paris, France, 1962.

[PS99] Ronan Pichon et Pascale Sébillot. Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In *6e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 99)*, Cargèse, France, 1999.

[PS00] Ronan Pichon et Pascale Sébillot. From Corpus to Lexicon: from Contexts to Semantic Features. In *Conference on Practical Applications in Language Corpora (PALC 99), Lodz Studies in Language*, Lodz, Pologne, 2000.

[PTL93] Fernando Pereira, Naftali Tishby, et Lillian Lee. Distributional Clustering of English Words. In *31st Annual Meeting of the Association for Computational Linguistics (ACL 93)*, Columbus, OH, EU, 1993.

[Pus95] James Pustejovsky. *The Generative Lexicon*. MIT press, Cambridge, MA, EU, 1995.

[Qui60] Willard van Orman Quine. *Word and Object*. MIT Press, Cambridge, MA, EU, 1960.

[Qui68] Ross Quillian. Semantic Memory. In Marvin L. Minsky, éditeur, *Semantic Information Processing*, pages 227–270. MIT Press, Cambridge, MA, EU, 1968.

[Ras72] François Rastier. Systématique des isotopies. In Algirdas Julien Greimas, éditeur, *Essais de sémiotique poétique*. Larousse, Paris, France, 1972.

[Ras89] François Rastier. *Sens et textualité*. Langue, linguistique, communication. Hachette, Paris, France, 1989.

[Ras96] François Rastier. *Sémantique Interprétative*. Presses Universitaires de France, Paris, France, 2e édition, 1996. 1ère édition en 1987.

[RCA94] François Rastier, Marc Cavazza, et Anne Abeillé. *Sémantique pour l'analyse : de la linguistique à l'informatique*. Masson, Paris, France, 1994.

[RS97] Ellen Riloff et Jessica Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. In *2nd Conference on Empirical Methods in Natural Language Processing*, Sommerset, NJ, EU, 1997.

[RS99] Ellen Riloff et Jessica Shepherd. A Corpus-based Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction. *Natural Language Engineering*, 5(2):147–156, 1999.

[RS02] Mathias Rossignol et Pascale Sébillot. Automatic Generation of Sets of Keywords for Theme Detection and Characterization. In *6e Journées internationales d'Analyse des Données Textuelles (JADT 2002)*, Saint-Malo, France, 2002.

[Sha48] Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 et 623–656, 1948.

[Sik72] Laurent Siklóssy. Natural Language Learning by Computer. In Herbert A. Simon et Laurent Siklòssy, éditeurs, *Representation and Meaning: Experiments with Information Processing Systems*, pages 288–328. Prentice-Hall, Englewood Cliffs, NJ, EU, 1972.

[SM99] Kai-Chung Siu et Helen M. Meng. Semi-Automatic Acquisition of Domain-Specific Semantic Structures. In *6th European Conference on Speech Communication and Technology (EuroSpeech 99)*, Budapest, Hongrie, 1999.

[Sme99] Alan F. Smeaton. Using NLP or NLP Resources for Information Retrieval Tasks. In Tomek Strzalkowski, éditeur, *Natural Language Information Retrieval*, pages 99–111. Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1999.

[Sow84] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA, EU, 1984.

[SSBM96] Gerard Salton, Amit Singhal, Chris Buckley, et Mandar Mitra. Automatic Text Decomposition Using Text Segments and Text Themes. In *7th ACM Conference on Hypertext (Hypertext 96)*, Washington D.C., EU, 1996.

[Tan97] Ludovic Tanguy. *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative*. Thèse de doctorat en informatique, Université de Rennes I, Rennes, France, 1997.

[THTSS01] Gökhan Tür, Dilek Z. Hakkani-Tür, Andreas Stolcke, et Elizabeth Shriberg. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*, 27(1):31–57, 2001.

[Vai99] Pascal Vaillant. *Sémiotique des langages d'icônes*. Bibliothèque de grammaire et de linguistique. Honoré Champion, Paris, France, 1999.

[Vos98] Piek Vossen, éditeur. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1998.

[War63] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236–245, 1963.

[Wea55] Warren Weaver. Translation. In William N. Locke et A. Donald Booth, éditeurs, *Machine Translation of Languages*, pages 15–23. John Wiley & Sons, New York, NJ, EU, 1955.

[Wid03] Dominic Widdows. A Mathematical Model for Context and Word-Meaning. In *4th International and Interdisciplinary Conference on Modelling and Using Context*, Stanford, CA, EU, 2003.

[Wie48] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. John Wiley & Sons, New York, EU, 1948.

[Zol68] Émile Zola. *Madeleine Férat*. Lacroix, Paris, France, 1868.

Table des figures

Résumé

Les lexiques sémantiques sont des ressources indispensables pour permettre à de nombreuses applications de traitement automatique des langues (résumé automatique, recherche d'information, traduction automatique, *etc.*) d'accéder au sens d'un texte. La question de la pertinence des informations présentes dans de tels lexiques est cruciale : le sens d'un mot comme *navet*, par exemple, varie considérablement selon que le texte étudié est consacré à la gastronomie ou à la cinématographie. Un moyen économique et linguistiquement motivé de disposer de descriptions de sens de mots réellement pertinentes est d'« apprendre » celles-ci à partir de l'utilisation effective des mots dans un *corpus*, collection de textes représentative d'un domaine d'expression.

Nous proposons pour tenter de répondre à ce défi une méthodologie d'acquisition automatique sur corpus d'informations lexicales sémantiques en trois étapes, fondée sur les principes linguistiques de la Sémantique interprétative de F. Rastier. Par analyse statistique et comparaison des modes d'emploi des mots, en utilisant des méthodes classiques ou originales, nous parvenons tout d'abord à rapprocher des mots appartenant à un même domaine (par exemple *donnée*, *transfert*, *réseau* pour les NTIC), puis des mots de sens similaires (*donnée* et *information*). Enfin, nous proposons une première méthode permettant la mise au jour de nuances fines marquant des distinctions de sens entre mots proches (*donnée* est plus « concret » que *information*), ce qui constitue un résultat encore inédit en acquisition automatique d'informations lexicales sémantiques.

Mots-clés

Traitement automatique des langues, apprentissage automatique, corpus, sémantique interprétative, mots-clés, thèmes, classes sémantiques, taxèmes, nuances sémantiques, sèmes spécifiques.

Abstract

Semantic lexicons are an essential resource to let many natural language processing applications (automatic summarization, information retrieval, automatic translation, *etc.*) penetrate the meaning of a text. The relevance of the information gathered by those lexicons raises a problematic question: the meaning of a word like *soap*, for example, varies considerably whether it is considered in a sanitary or televisual context. A linguistically motivated and cost-effective way of building semantic lexicons precisely adapted to a certain domain of expression consists in "learning" word meanings from their actual usage as observed in a representative collection of texts, or *corpus*.

To answer this challenge, we propose in this document a three-stage methodology for the automatic acquisition of lexical semantic information from texts, based on the linguistic principles of F. Rastier's Interpretative semantics. Thanks to a statistical analysis of word uses, employing both classical and novel methods, we first manage to bring together words belonging to a same domain (for example *data*, *transfer*, *network* for IT), then to build classes of words having a similar meaning (*data* and *information*). We finally propose a first method to put to light fine-grained meaning distinctions between close words (*data* is more "concrete" than *information*), thus reaching a level of meaning refinement never before attained, to our knowledge, by automatic means.

Keywords

Natural language processing, machine learning, corpus, interpretative semantics, keywords, topics, semantic classes, taxemes, semantic nuances, specific semes.