Construction et interrogation de la structure informationnelle d'une base documentaire en français

Bernard Jacquemin

Université Paris III Sorbonne Nouvelle

ILPGA
Institut de Phonétique et de Linguistique Générales et Appliquées

Thèse de doctorat en Sciences du Langage
Linguistique et Informatique

Construction et interrogation de la structure informationnelle d'une base documentaire en français

Bernard JACQUEMIN

Soutenue publiquement le 8 décembre 2003 devant le jury composé de :

M. Henri BÉJOINT	Rapporteur
M. Benoît HABERT	Directeur
M. Claude ROUX	Examinateur
M. André SALEM	Président
Mme Frédérique SEGOND	Examinatrice
Mme Monique SLODZIAN	Rapporteur

Merci à tous !

Si j'en arrive maintenant au terme de ces années de recherche et de rédaction, c'est grâce à vous tous, à qui je veux dire un grand merci.

À Benoît Habert, mon directeur de thèse, qui m'a dirigé tout en me laissant une grande autonomie, et dont les remarques toujours pertinentes et les questions m'ont poussé toujours plus loin dans ma réflexion et vers mon objectif.

À Frédérique Segond qui m'a ouvert les portes de la recherche en même temps que celles de Xerox, m'a poussé à entreprendre cette thèse et, en m'apprenant un métier, m'a aidé à ne pas renoncer.

À Claude Roux dont les connaissances et les compétences m'ont accompagné pas à pas tout au long de cette thèse. Sans lui, elle n'existerait pas.

Au professeur André Salem à qui échoit la charge de présider mon jury.

Aux professeurs Monique Slodzian et Henri Béjoint qui ont eu la grande gentillesse de mettre leur savoir au service de mon travail en acceptant d'en être les rapporteurs.

Au professeur Christiane Fellbaum pour sa gentillesse, la grande générosité dont elle a fait preuve à mon égard et aussi, malgré l'impossibilité dans laquelle elle était de prendre part à mon jury, pour avoir accepté de parcourir ma thèse et de l'avoir enrichie de ses remarques et suggestions.

Aux professeurs Sylviane Granger et Guy Deville, qui m'ont initié à la linguistique computationnelle et, ce faisant, m'ont fait découvrir tout un monde que je ne connaissais pas.

À Jean-Luc Manguin, du CRISCO à Caen, qui m'a fourni aimablement les sources du *Dictionnaire des synonymes de la langue française* de René Bailly.

Aux différentes personnes qui ont consacré du temps et de l'énergie à relire les épreuves successives de mon manuscrit : Marie-Hélène Corréard, Veronika Lux, Caroline Brun, Karen Fort, Nuria Gala-Barbaste, Salah Aït-Mokhtar.

Aux autres membres des équipes auxquelles j'ai appartenu au sein de XRCE, qui m'ont bien souvent éclairé de leur savoir, et plus particulièrement Anne

Schiller, Caroline Hagège, Ken Beesley, Éric Gaussier, François Trouilleux, Nicola Cancedda, Aaron Kaplan, Hervé Déjean.

À mes utilisateurs « spécialistes », cités, et « naïfs », Damián Arregui, Franck Guingne, Florent Nicart et Frédéric Roussey, qui ont rendu mon évaluation possible.

À mes prédécesseurs en thèse à Xerox, pour toutes ces conversations enrichissantes, ou pas : Laurent Griot, Mathieu Mangeot, Julien Quint, Sylvain Pogodalla.

À tous mes collègues, qui sont aussi devenu des amis au cours de ces années, et particulièrement Denys (qui tient au « y »), Nicolas (et Christel), Pico, Nate, Manfred bien sûr, Christine (et Michaël), Séverine et tant d'autres.

À Jean-Pierre Chanod, dans l'équipe de qui j'ai été versé lorsque Frédérique a été appelée à d'autres responsabilités.

À ma famille dont le soutien est indéfectible.

À vous tous donc, j'exprime ici ma gratitude la plus profonde.

Indications typographiques

La police généralement utilisée dans cette thèse est généralement du Roman standard sous LaTeX2e, avec les adaptations typographiques dues à la structure du texte, notamment les variations de taille et de style pour les titres, les notes de bas de page, les légendes de figures, etc. Toutefois, pour rendre le texte plus lisible, nous avons utilisé les possibilités offertes par l'éditeur pour distinguer certains énoncés ou fragments d'énoncés.

Les **caractères gras** peuvent indiquer soit la première apparition dans le texte d'un terme traité dans le glossaire, soit insister sur les points remarquables des exemples présentés.

Les *caractères penchés* permettent de différencier les mots qui appartiennent à une langue étrangère, principalement l'anglais ou le latin.

Les *caractères italiques* sont utilisés pour signaler les exemples ou les extraits d'exemples dans le corps du texte, ainsi que les titres d'ouvrages ou les noms de logiciels.

Les `caractères de machine à écrire` expriment dans le texte ou dans les figures et tableaux les résultats d'un traitement automatique tels qu'ils se présentent à l'écran d'un ordinateur.

Les caractères sans sérif distinguent les commentaires sur les exemples dans les figures et les tableaux.

Table des matières

Table des figures

Liste des tableaux

Introduction

La société de l'information dans laquelle nous vivons a fait des documents sous forme électronique et de la maîtrise de l'information qu'ils contiennent un enjeu majeur dans des secteurs aussi variés que la politique, l'économie, la culture, la défense... Toutefois, l'accès à cette information est rendu malaisé par deux facteurs. D'abord, ces documents électroniques sont actuellement si nombreux qu'aucun humain ne peut en avoir une idée claire, et encore moins en maîtriser le contenu – d'autant plus que le volume des textes continue d'augmenter de plus en plus vite. Ensuite, l'absence presque généralisée d'organisation au sein de cette profusion de textes ne facilite pas l'accès à une information précise.

Le domaine de la **gestion de l'information** textuelle, qui se penche précisément sur cette problématique, tente d'y remédier par des approches automatiques. Les différentes disciplines qui le composent se consacrent donc à élaborer des stratégies permettant de repérer la présence d'un ou plusieurs éléments d'information à l'intérieur de textes. Les disciplines les plus représentatives – car elles sont également les plus exigeantes – sont celles de **question-réponse** et d'**extraction d'information**. Toutes deux cherchent en effet l'identification précise d'une information dans les textes, soit parce qu'elle répond à une question posée, soit parce qu'elle correspond à un type informationnel prédéfini. La détermination précise de l'information recherchée constitue la grande difficulté de ces tâches.

Du fait de la grande activité du domaine, les méthodes mises en œuvre dans ce cadre sont multiples. Toutefois, les systèmes existants ont en commun une approche centrée sur l'étude de la nature de l'information désirée, c'est-à-dire la question en question-réponse et le type informationnel en extraction d'information. Ils effectuent ensuite une comparaison de cette information avec le contenu des textes, puis un calcul de distance entre information désirée et information obtenue. De ce fait, l'examen des documents est secondaire. D'autre part, la supériorité des systèmes qui intègrent les éléments d'analyse linguistique les plus complexes a été constatée [Ferret et al., 2002a]. Cependant, aucune approche n'a poussé le raisonnement jusqu'à se fonder entièrement sur des outils linguistiques.

Inscrite dans le cadre du **TAL**, la recherche que nous abordons dans cette thèse entend se pencher sur la problématique de la gestion de l'information à partir d'un angle nouveau. En effet, comme la désorganisation des documents électronique est un obstacle à l'accessibilité à l'information, notre idée est de construire à partir d'une base textuelle [1] la structure correspondant à l'information qui y est contenue. Cette structure constituerait de ce fait un point d'accès aisé à toute information désirée.

La construction de cette **structure informationnelle** doit se plier à deux impératifs. Tout d'abord, la profusion des textes demeurant un problème auquel le domaine doit faire face, il est essentiel que la démarche soit automatique et que l'intervention humaine la plus petite possible soit requise. Ensuite, puisque les processus linguistiques ont établi leur intérêt et que notre démarche se veut axée sur le TAL, nous avons décidé de privilégier les approches linguistiques des textes au détriment des méthodes statistiques et quantitatives. Ainsi l'examen des documents dont il nous faut construire la structure informationnelle doit-il reposer sur des outils d'analyse textuelle basés sur la grammaire de la langue et sur des dictionnaires. Les énoncés des documents, plus larges que ceux des requêtes (questions ou type informationnel désiré) conviennent d'ailleurs souvent mieux à une analyse linguistique.

La structure informationnelle se doit de contenir les éléments d'information contenus dans les textes (les mots), d'en identifier la nature (le sens des mots) ainsi que les rapports qu'ils entretiennent entre eux (les relations syntaxiques). Des analyses morphologique et syntaxique doivent donc être effectuées, ainsi qu'une **désambiguïsation sémantique**. Les résultats de ces opérations constituent le corps de la structure, qui permettent d'obtenir les fragments auxquels ils correspondent. Un dictionnaire, exploité par la désambiguïsation sémantique, sert de référent au niveau du sens.

La richesse de la structure est le garant de son utilité. Elle doit contenir l'ensemble de l'information présente dans les textes auxquels elle fait référence. Mais elle doit aussi comporter, pour chaque information, le plus grand nombre possible de leurs réalisations lexico-sémantiques ou syntaxiques, afin que chaque information puisse être atteinte quelle que soit la forme de la requête qui y fait référence. Là encore, les disciplines de question-réponse et d'extraction d'information agissent de manière similaire, l'une effectuant l'expansion de la requête, l'autre constituant des listes de patrons ou de tableaux correspondant au type informationnel recherché.

Toutefois, nous pensons que l'**enrichissement** apporté à la structure doit correspondre aux éléments d'information collectés, à leur nature et aux rapports qu'il y a entre eux. Dès lors, l'enrichissement devra permettre d'ajouter à la structure des données qui correspondent à des énoncés différents de ceux des textes, mais qui possèdent la même signification. Les données lexicales

[1] Les textes qui constituent cette base documentaire sont en français.

récoltées dans des dictionnaires, parfois modifiés de manière à en utiliser uniquement les éléments assignés au sens voulu, permettent de nombreuses adjonctions à la structure, et ouvrent l'accès à l'information qu'elle contient.

Cet accès peut s'effectuer de diverses manières, dont la plus simple est par mots-clefs. Toutefois, dans la perspective du TAL, il nous a paru intéressant de tenter d'accéder à l'information par requête en langage naturel. Le principe sera donc de constituer une structure locale à la requête et de la comparer à la structure informationnelle enrichie de la base documentaire interrogée. Les traitements appliqués à la requête sont semblables à ceux des documents, excepté la désambiguïsation sémantique – le contexte étant généralement trop exigu pour permettre son application – et l'enrichissement. De plus, certaines règles de grammaire permettent d'identifier dans l'énoncé l'objet de la requête, et d'éliminer les éléments interrogatifs de la structure locale. La comparaison entre la structure locale de la question et celle des documents peut alors avoir lieu, et les critères de correspondance entre la question et les fragments de texte susceptibles d'en constituer la réponse ont la possibilité d'être plus ou moins rigoureux. Les différents types d'enrichissement peuvent même être sélectionnés ou rejeter individuellement.

Enfin, pour tester la validité de notre approche, il est capital de disposer d'une base documentaire en langue française, de taille suffisante pour constituer un corpus de travail et un corpus d'évaluation cohérents et homogènes, qui permettent de tester non seulement la capacité de l'approche à apporter une réponse correcte à une question posée (question-réponse), mais aussi à en trouver toutes les réponses (extraction d'information).

Or notre thèse est effectuée dans le cadre d'une convention **CIFRE** entre le laboratoire **ILPGA** (Institut de Linguistique et Phonétique Générales et Appliquées, Université de la Sorbonne Nouvelle – Paris III) et XRCE (Xerox Research Centre Europe) à Grenoble. Ce dernier laboratoire dispose, dans le cadre d'un projet appelé CIRCE, d'une version électronique du texte de l'Encyclopédie Multimédia Hachette[2]. Il s'agit d'un dictionnaire encyclopédique en français. Cet ouvrage présente divers avantages. Les documents sont relativement vierges d'erreurs orthographiques. Ils portent sur des sujets variés souvent susceptibles de recouper l'information d'autres documents tout en ne portant jamais sur le même sujet. Ils sont composés de telle manière qu'il est facilement possible de constituer des ensembles de textes comportant une information cohérente. Enfin, tous les documents présentent une constitution interne comportant un « cartouche » de texte structuré sémantiquement, composé du titre de l'article et de certaines informations résumées et formatées.

Le texte de ce dictionnaire encyclopédique se présente sous la forme d'une multitude de fichiers informatiques encodés en un langage balisé bap-

[2]C'est la version 2 000 de cette encyclopédie électronique dont dispose le projet CIRCE.

tisé **XML** (*eXtensible Markup Language*). Il s'agit d'un langage-outil qui permet d'ajouter à un document différentes informations, comme une présentation particulière – de manière similaire au langage **HTML** d'Internet – ou des indications d'ordre sémantique. Chacun des fichiers correspond à un article de l'encyclopédie, la réalité traitée dans l'article en constituant le titre. Cette encyclopédie est donc constituée comme un dictionnaire et elle est alphabétiquement ordonnée. Elle contient d'ailleurs un dictionnaire de langue général composé de trente-cinq mille entrées [3] environ sur les quelque septante-cinq mille qui composent l'entièreté de l'ouvrage. Cette partie lexicale ne concerne pas la constitution d'une base de textes informationnelle étant donné que les énoncés qu'elle contient relèvent par nature du domaine spécialisé de la lexicographie.

En tant qu'encyclopédie générale, les sujets abordés sont aussi divers que l'histoire, l'économie, la littérature ou la chimie. La longueur des articles est elle aussi extrêmement variable, de quelques lignes à plusieurs pages [4] et le type des sujets traités, bien que portant le plus souvent sur des personnes physiques, peuvent également varier et concerner des pays, villes, œuvres, faits de société ou des événements. Cette variété concourt à rendre ce type de document propre à servir de base documentaire de test pour notre méthode de désambiguïsation sémantique.

Nous l'avons dit plus haut, le texte de l'encyclopédie n'est pas brut, mais il est enrichi de balises et de codes XML (*cf.* figure 1 page ci-contre). Ainsi, les articles de l'encyclopédie encodent-ils habituellement l'entrée de l'article comme `Sommaire` et comme `Titre` et les divers paragraphes de son corps comme `Parency`. Les différents types d'articles sont distingués entre eux, les entrées du dictionnaire de langue portant l'indication `Entree.lang`, tandis que celles propres à l'encyclopédie sont signalées par `Entree.ency`, et les légendes des illustrations comportent `Entree.leg`. Sémantiquement, le titre de l'article contient le `nom` et le `prenom` des personnages, les autres sujets qui ne concernent pas une personne sont quant à eux étiquetés simplement `nom`. D'un point de vue purement typographique, on peut voir dans l'exemple que le `Nom` (et le `Prenom` dans le cas d'un personnage) du `Sommaire` et du `Titre` est en caractères gras (`bold`). On a encore des indications de `Lieu` et de `Date` pour la `naissance` et la `mort` des personnages. Dans le corps même de l'article, outre la séparation en paragraphes dont nous avons parlé plus haut (`Parency`), les intitulés des œuvres, titres de journaux et toute entité à laquelle un article est susceptible d'être consacré portent l'étiquette

[3] Le *Petit Robert* contient soixante mille entrées environ et le *Petit Larousse* cinquante-neuf mille.

[4] La taille du texte peut être exprimée de différentes manières. Lorsque nous parlons de volume, il s'agit de l'occupation du texte électronique sur un disque et dans ce cas, le balisage XML fait partie de ce volume. Au contraire, lorsque nous mentionnons sa longueur, nous sous-entendons qu'il s'agit d'une version lisible, épurée de tout balisage ou codage XML.

`Ref`. Les formules mathématiques sont étiquetées `Formule`. Étant donné que les articles de langue ne nous concernent pas, nous ne mentionnons pas les étiquettes qui leur sont propres.

```
<Entree.ency>
  <Sommaire type="T" id="g0004130.1">
    <Nom>
      <Format typo="bold">Gunter</Format>
    </Nom>(
    <Prenom>
      <Format typo="bold">Edmund</Format>
    </Prenom>)
  </Sommaire>
  <Titre type="T">
    <Nom>
    [...]
    </Prenom>)
  </Titre>
  <Resume>Astronome anglais</Resume>
  <Etatcv>(
    <Lieu type="naissance">dans le Hertfordshire</Lieu>
    <Date type="naissance">, 1581</Date>
    <Lieu type="mort">&agrave; Londres</Lieu>
    <Date type="mort">, 1626</Date>)
  </Etatcv>
  <Parency>
    Il fut l&#39;auteur de tables trigonom&eacute;triques &agrave;
    7 d&eacute;cimales et des notions de cosinus et de sinus; il
    inventa diff&eacute;rents instruments d&#39;astronomie et de
    navigation (r&egrave;gle de Gunter).
  </Parency>
</Entree.ency>
```

FIG. 1 – Exemple d'entrée de l'*Encyclopédie Hachette* au format XML.

C'est donc sur base d'un corpus prélevé dans les différents articles de ce dictionnaire encyclopédique que les tests sont effectués au long de notre recherche et de la construction du système de construction de structure informationnelle. Un autre corpus est construit pour la phase d'évaluation du système complet.

Le premier chapitre de cette thèse se penche sur les conférences qui font référence dans le domaine visé, et plus particulièrement **MUC**, qui a défini les objectifs de l'extraction d'information, et **TREC**, qui a fait de même pour la tâche de question-réponse entre autre. Ce chapitre est également consacré à l'examen de diverses approches de ces deux disciplines, aux besoins qu'elles affichent et aux obstacles qu'elles rencontrent.

Le second chapitre est consacré à l'étude des outils linguistiques utilisés pour l'analyse textuelle, et plus particulièrement à la problématique de la désambiguïsation sémantique qui est essentielle dans l'identification de la

nature de l'information rencontrée. À ce titre, elle sert aussi de référence pour l'enrichissement de la structure informationnelle.

Le troisième chapitre a pour objet la description et le choix d'un dictionnaire de référence qui sera utilisé par le système de désambiguïsation sémantique. D'autres ressources lexicales sont également sélectionnées qui auront leur utilité lors de la phase d'enrichissement de la structure informationnelle.

Dans le chapitre quatre, nous décrivons les traitements que nous appliquons aux différentes ressources lexicales pour en modifier certains aspects qui ne conviennent pas à notre méthodologie et pour les rendre compatibles les uns avec les autres.

Le cinquième chapitre se consacre à la construction de la structure informationnelle, d'abord à l'aide des résultats obtenus à travers l'analyse linguistique des documents, ensuite grâce aux différents types d'enrichissement mis en œuvre, qui sont décrits eux aussi.

La méthode d'interrogation de la structure informationnelle en langage naturel fait l'objet du sixième chapitre. Les traitements particuliers appliqués à la question ainsi que les différents niveaux de comparaison entre la question et la réponse y sont indiqués.

Le septième chapitre sanctionne la méthode élaborée par une évaluation. Un protocole d'évaluation est d'abord exposé, puis les résultats sont présentés et commentés.

Enfin, nous discutons des apports et des défauts de notre méthodologie, et nous en présentons les perspectives.

Chapitre 1

Le problème de la gestion de l'information

1.1 Introduction

Notre thème de recherche a pour objet d'élaborer une méthode de construction de structure informationnelle à partir d'une base documentaire. Cette structure doit être capable de fournir un accès à la réponse d'une question posée relative au contenu des documents. De nombreuses recherches ont été menées depuis les débuts du traitement automatique des documents qui poursuivent un objectif semblable : trouver une information précise dans des textes. Notamment, les techniques d'extraction d'information et de question-réponse appartiennent aux deux disciplines les plus exigeantes en ce qui concerne l'identification de l'information recherchée.

Malgré leurs spécificités, ces deux disciplines possèdent en commun différentes caractéristiques. Premièrement, toutes deux cherchent à identifier l'information recherchée grâce à des répertoires – préalablement constitués ou non – capables de repérer toutes les formes sous lesquelles l'information recherchée peut se présenter. L'extraction d'information constitue donc des ensembles de patrons ou de cadres capables de retrouver un type informationnel sous un grand nombre d'aspects. De leur côté, les approches de question-réponse travaillent en expansion de la requête proposée et fournissent pour les données qui s'y trouvent un maximum d'**actualisations** différentes.

Un second point commun entre les deux disciplines concerne leur intérêt croissant pour l'analyse textuelle et les approches linguistiques. En effet, l'identification des mots permet de préciser la nature de l'information recherchée dans les textes et aussi celle de l'information présente dans la question ou le type informationnel. Les aspects statistiques dominent toutefois dans

la plupart des approches.

Ce chapitre s'intéresse à la conférence MUC spécialisée en extraction d'information et aux méthodologies qui y ont été présentées. Il se penche ensuite sur la campagne d'évaluation de TREC consacrée à la discipline de question-réponse et sur certains systèmes qui y ont concouru. Il s'agit en effet de connaître les besoins des domaines auxquels nous proposons notre méthodologie, et d'identifier également les points forts des méthodologies proposées comme les difficultés auxquelles elle doivent faire face.

1.2 Analyser un texte pour en extraire l'information

1.2.1 Un bref historique

C'est après la seconde guerre mondiale et plus précisément avec l'accroissement des tensions Est-Ouest que les systèmes d'extraction automatique d'information ont fait leur apparition. Les militaires désiraient en effet collecter un maximum d'informations géopolitiques dans les documents publics ou secrets qu'ils avaient à leur disposition, mais dont la masse ne permettait pas une lecture humaine complète. La fin de la guerre froide puis la chute du bloc communiste a mis fin à cette mainmise militaire sur le domaine de l'extraction d'information, avec l'organisation, dès 1987, de la conférence MUC (*Message Understanding Conference*).

Cette conférence, dont le thème est la compréhension automatique de textes et l'extraction de leur information, a progressivement créé un consensus sur la définition de ces domaines [Appelt, 1999]. Elle a formellement défini la notion d'extraction d'information et sert actuellement de référence dans cette matière. Les publications qui lui sont attachées constituent l'état de la recherche du domaine. Elle est organisée selon le principe de la compétition entre les systèmes participants qui doivent donc se plier à ses critères de travail et d'évaluation [Chinchor, 1992]. Notamment, il s'agit de présenter un système dont le résultat soit une structure hiérarchique d'attributs-valeurs couramment appelée **formulaire** (*template*) et d'accepter l'évaluation et la publication des résultats dudit système. Cette évaluation s'effectue sur la base de corpus de taille moyenne pour obtenir les formulaires de référence à l'idéal. Ces textes sont au nombre de 1 300, contenant environ 400 000 mots pour un vocabulaire de 18 000 mots. Les textes contiennent 12 phrases composées en moyenne de 27 mots [Chinchor et al., 1994].

Est considéré comme **correct** (C) un formulaire extrait dont la valeur est conforme à celle du formulaire de référence correspondant. Si cette valeur ne correspond pas, le formulaire est **incorrect** (I). Dans le cas où un attribut

est présent dans le formulaire extrait mais qu'il est différent de l'attribut du formulaire de référence correspondant, il y a **surgénération** (O). À l'aide de ces résultats, il est possible de déterminer la **précision** (P) et le **rappel** (*recall*) (R) du système :

$$précision = \frac{C}{C+I+O}$$

$$rappel = \frac{C}{P}$$

On appelle **F-mesure** (*F-measure*) un résultat statistique qui découle de la combinaison de la précision et du rappel, pondéré par un paramètre β dont la variation à partir de 1.0 détermine si le rappel ou la précision est de plus de poids :

$$F = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

La précision, le rappel et la F-mesure sont les mesures les plus fréquentes qui permettent d'évaluer les capacités d'un système d'extraction d'information. La précision indique la proportion des réponses correctes dans l'ensemble des réponses apportées par le système. Le rappel évalue le rapport entre le nombre de réponses correctes apportées par le système et l'ensemble des réponses correctes présentes dans la base interrogée [Salton et McFill, 1983]. La F-mesure est une mesure qui calcule l'harmonisation pondérée entre la précision et le rappel pour l'évaluation des systèmes. La pondération est habituellement de β=1.0 pour harmoniser l'importance conférée à la précision et au rappel [Makhoul et al., 1999], mais il est parfois utile de montrer le comportement des applications en privilégiant également la précision (β=0.5) et le rappel (β=2.0) [Lewis, 1992].

Par ailleurs, la conférence MUC a défini deux modes d'évaluation des systèmes qui lui sont soumis. L'un de ces modes met l'accent sur la précision et détermine la perfection du système par des résultats de 100% de rappel pour 100% de précision. Le second mode privilégie la pureté des réponses et établit l'excellence par des résultats de 100% de rappel pour 0% de surgénération.

Des diverses applications présentées pour concourir, les conférences MUC successives ont dégagé différentes tâches générales vers lesquelles ces systèmes tendaient. En fonction de leur objectif – la plupart de ces systèmes ne traitent pas l'intégralité des tâches recensées par MUC –, les systèmes sont évalués en fonction de leur capacité à résoudre ces tâches :

- Reconnaissance d'entités nommées. Relevé des noms propres et identification de l'entité qu'ils recouvrent à l'intérieur du texte (personnes, sociétés, lieux...).
- Construction de formulaires élémentaires. L'identification des entités nommées peut ne pas suffire. En effet, il y a des cas où la même entité

peut se présenter sous plusieurs formes différentes, qu'il faut identifier l'une à l'autre, ou inversement deux entités distinctes peuvent être désignées de la même manière, et il s'agit alors de distinguer ces individus (*Joconde*, *Mona Lisa*). Certains attributs peuvent être pris en compte (titre, nationalité...).

- Mise en relation des formulaires. Construction de relations entre les éléments des formulaires identifiés lors de la construction de formulaires élémentaires (par exemple une relation *invention* entre un chercheur et sa découverte). Il s'agit d'une tâche capitale en extraction d'information mais elle ne peut être entièrement réalisée car le nombre de relations entre les différents éléments du monde est infini.
- Traitement de la **coréférence**. Identification des expressions faisant référence à une même entité et mise en rapport de ces expressions équivalentes.
- Élaboration d'un scénario global. Mise en commun des éléments d'information collectés précédemment en une seule structure, le formulaire (événement, entité, relation, attribut...).

Suite au succès rencontré par MUC, d'autres conférences ont rapidement vu le jour, qui abordent des sujets semblables ou connexes. Notamment, depuis sa première édition en 1992, la compétition TREC a pris une place prépondérante dans le domaine de la gestion de l'information textuelle [1] [Harman, 1992]. Moins ambitieuse que MUC, TREC a pour objectif non pas la compréhension de textes, mais l'identification dans des documents d'éléments informationnels qui composent une information recherchée, préalablement définie ou non. Cette définition d'un objectif moins strict que celui de MUC permet d'envisager plusieurs facettes du problème.

Ces différentes perspectives s'actualisent autour de plusieurs tâches, présentes dès l'origine de la conférence ou qui ont vu le jour au cours des différentes éditions.

Les systèmes que nous décrivons dans ce chapitre viennent à la suite de l'initiative MUC. Ils possèdent en commun une approche automatisée de l'apprentissage de structures textuelles, sémantiques ou non, qui leur permettent d'identifier l'information pertinente dans les textes, et donc de constituer une structure sémantique plus ou moins complète des textes utilisés. À ce titre, ils constituent un aperçu des méthodes de construction automatique d'un **dictionnaire sémantique** local en extraction d'information. De ce fait, leur domaine d'application est généralement spécialisé. Notre démarche, bien que visant une compréhension plus globale des textes, s'appuie sur certains processus similaires à ceux que ces systèmes mettent en œuvre.

[1] Ce n'est pas seulement l'information textuelle qui intéresse TREC, mais aussi l'image, le son, etc. Toutefois, cet aspect seul nous intéresse dans le cadre de cette thèse.

1.2.2 Quelques notions

Types de textes

Il existe trois classes de systèmes d'extraction d'information, selon le type de texte que le système est capable de gérer, même si certaines approches peuvent être polyvalentes [Soderland, 1999]. Le **texte libre** correspond au langage écrit selon les règles grammaticales en vigueur dans la langue utilisée mais sans autre contrainte. Un article de journal ou le texte d'un roman est généralement libre, comme le texte d'une encyclopédie :

> « Le génie de Victor Hugo est partout, dans tous les genres littéraires qu'il bouleverse et s'approprie – du roman au drame, de la poésie à la critique, de l'ode au pamphlet – comme dans cette facilité à se faire l'acteur des grands rôles publics. » [2]

Le **texte structuré** répond à des règles très strictes et l'information dont il est porteur est régulière dans sa nature, sa présence, sa position. L'information cataloguée (date, œuvre, genre) de ce tableau

1829	Marion Delorme	Drame
1830	Hernani	Drame
1831	Notre-Dame de Paris	Roman

correspond au texte suivant [3], dont la structure est concrétisée par des balises :

```
<tr class="cell2" valign="top">
  <td align="center"> 1829 </td>
  <td align="left"> Marion Delorme </td>
  <td align="left"> Drame </td>
</tr>
<tr class="cell1" valign="top">
  <td align="center"> 1830 </td>
  <td align="left"> Hernani </td>
  <td align="left"> Drame </td> </tr>
<tr class="cell2" valign="top">
  <td align="center"> 1831 </td>
  <td align="left"> Notre-Dame de Paris </td>
  <td align="left"> Roman </td>
</tr>
```

Enfin, le **texte semi-structuré** suit rarement la grammaire de la langue et se présente sous un style plus ou moins télégraphique, sans règle rigide ni sans forme prédéfinie. Il est en général porteur d'une information utilitaire.

[2] *Encyclopédie Hachette Multimédia*, sous l'entrée *Victor Hugo*.
[3] *Encyclopédie Hachette Multimédia*, extrait de la *Bibliographie de Victor Hugo*.

Le texte que nous devons traiter dans le cadre de cette recherche est à l'intersection du texte libre, puisque le contenu est en général grammatical et que l'information qu'il contient est imprévisible, et du texte semi-structuré, étant donné qu'il s'agit d'un dictionnaire encyclopédique qui répond à des règles de présentation dans l'intitulé de ses articles et qu'une structure XML sous-jacente en identifie certains éléments.

Types de formulaires

La normalisation des résultats de l'extraction d'information au sein de la conférence MUC demande une structure attribut-valeur appelée formulaire dans le cadre de la conférence. En réalité, le sens de ce vocable varie selon les approches, et cette structure porte tantôt le nom de « signature » [Riloff et Lorenzen, 1999], tantôt celui de « cadre » (*frame*) [Soderland, 1999] ou de « patron » (*pattern*) [Kim et Moldovan, 1993]. Il reste toutefois que cette structure, quel que soit le nom que l'on lui donne, contient toujours au moins un « emplacement » ou une « case » (*slot*) destiné à recevoir un élément d'information d'une catégorie sémantique déterminée.

```
Input : Mr. Adams, former president of X Corp., was named
CEO of Y Inc.
Succession event
        PersonOut :  Adams
        Post :       president
        Org :        X Corp.

Succession event
        PersonIn :   Adams
        Post :       CEO
        Org :        Y Inc.
```

FIG. 1.1 – Exemple de structure d'extraction combinée.

Or là où certains systèmes peuvent mettre en rapport, au sein d'une même structure, différents emplacements spécifiques ou non d'une même catégorie sémantique (**structure d'extraction combinée**, *multi-slot*, *cf.* figure 1.1), d'autres en sont incapables (**structure d'extraction simple**, *single-slot*, *cf.* figure 1.2 page ci-contre). Cela peut dans certains cas se révéler handicapant pour le système lorsque les relations entre différents emplacements peuvent les distinguer. L'exemple de la figure 1.2 page suivante montre bien le problème qui peut découler d'une structure d'extraction simple : on ne peut déterminer d'après la structure le poste quitté ou le nouveau poste, ni la société à laquelle ces postes sont reliés.

```
Input : Mr. Adams, former president of X Corp., was named
CEO of Y Inc.
        PersonOut :  Adams
        PersonIn :   Adams
        Post :       president
        Post :       CEO
        Org :        X Corp.
        Org :        Y Inc.
```

FIG. 1.2 – Exemple de structure d'extraction simple.

Toutefois, les structures d'extraction simple s'avèrent adéquates dans les cas où une seule information de chaque catégorie est présente dans un texte distinct. Ce présupposé limite donc ce type d'extraction aux textes structurés et, éventuellement, aux semi-structurés, où l'information est plus prévisible qu'en texte libre.

1.2.3 Approches visant les documents structurés et semi-structurés

On pourrait s'étonner de notre intérêt pour les approches des textes structurés ou semi-structurés. En effet, nous travaillons sur une encyclopédie dont le texte est libre. Toutefois, puisqu'une certaine quantité d'information est balisée – et donc partiellement structurée – et identifiée selon une définition de type de document (*Document Type Definition*, **DTD**) XML, il nous a semblé pertinent de ne pas rejeter *a priori* les méthodes qui permettent de traiter le texte structuré.

L'analyse des pages web par *Wrapper Induction*

Avec *Wrapper Induction* [Kushmerick et al., 1997], nous nous intéressons à un système d'extraction d'information destiné à traiter des données structurées dans des tableaux et à en identifier les éléments en conservant leur cohérence avec les autres composants d'un même tableau. Ce système est typiquement voué à gérer des pages Web. De ce fait, son domaine d'application peut s'ouvrir à une extraction combinée autant qu'à une extraction simple, pour autant que les textes qui lui sont présentés soient résolument structurés.

Ce système repose sur l'utilisation de ***wrappers***, c'est-à-dire de procédures logicielles spécifiques à un type de structure de ressource informationnelle, et qui traduisent la réponse à une requête donnée en un nouveau canevas d'information simple ou combinée selon la structure de base du document

et le sujet de l'information sélectionné. Ce canevas identifie et, le cas échéant, combine les différents éléments d'une information de sujet prédéterminé.

Toutefois les *wrappers* sont généralement construits manuellement. L'originalité de [Kushmerick et al., 1997] réside donc dans la proposition d'une méthode inductive qui permet d'apprendre automatiquement l'organisation de documents dont l'information, combinée ou non, a été préalablement étiquetée. De la sorte, une information pourra être extraite de documents présentant une information de même nature et une structure de même type que celles des documents qui ont servi à générer le *wrapper*.

Ces *wrappers* reposent sur la génération de règles **HLRT** (pour *Head Left Right Tail*), dont le principe consiste à identifier les bornes gauche et droite de chaque élément d'information. La détermination de la structure d'en-tête et de fin de page permet en outre de ne pas limiter le nombre des canevas informatifs dans la page. En effet, aussi longtemps que la structure de fin de page n'est pas détectée, un canevas informationnel complet est susceptible d'être suivi par un autre, délimité lui aussi par les mêmes bornes.

```
<HTML>
  <TITLE>Some Country Codes</TITLE>
  <BODY>
      <B>Some Country Codes</B>
      <P>
      <B>Congo</B> <I>242</I><BR>
      <B>Egypt</B> <I>20</I><BR>
      <B>Belize</B> <I>501</I><BR>
      <B>Spain</B> <I>34</I><BR>
      <HR><B>End</B>
  </BODY>
</HTML>
```

FIG. 1.3 – Exemple de texte structuré proposé au traitement d'un *wrapper*.

L'exemple 1.4 page ci-contre illustre bien le fonctionnement du *wrapper*. Il parcourt le texte jusqu'à la borne de fin d'en-tête `<P>` mais n'en retient rien. Une fois cette borne dépassée, il recherche une borne d'information gauche (`<B>` ou `<I>`), à partir de laquelle il extrait l'information jusqu'à la borne droite correspondante suivante (`</B>` ou `</I>`), et imprime l'information extraite. Il reprend ensuite cette opération de recherche - extraction - impression jusqu'à la première occurrence de la borne initiale de fin de page (`<HR>`) qui marque la fin de la zone informative pertinente de la page et arrête le *wrapper* pour cette page. On obtient donc la liste des éléments

Éliminer texte jusqu'après première occurrence de <P>
Aussi longtemps que prochain <B> est avant prochain <HR>
  pour toute borne d'info gauche ou droite [<B>, <I> ou </B>, </I>]
    chercher occurrence suivante d'une borne gauche
    extraire info comprise entre cette borne et borne droite suivante
  retourner info extraites

FIG. 1.4 – Algorithme d'extraction pour le *wrapper* correspondant au texte structuré de la figure 1.3.

```
Congo
242
Egypt
20
Belize
501
Spain
34
```

FIG. 1.5 – Résultat de l'extraction d'information du texte présenté à la figure 1.3 par l'algorithme du *wrapper* de l'exemple 1.4.

d'informations compris entre les bornes `<B>` et `</B>`, `<I>` et `</I>` du corps de la page, à l'exclusion de l'en-tête et du bas de la page.

Cette méthode présente l'avantage d'une grande fiabilité : en effet, lorsque la structure du document traité est commune avec celle du corpus d'apprentissage, elle extrait sans difficulté toute l'information jugée pertinente lors de l'inférence du *wrapper*. D'autre part, son fonctionnement par bornes basées sur les balises d'un langage structuré s'applique bien à la partie la plus structurée des articles de l'encyclopédie que nous avons pour tâche de traiter, d'autant plus que l'information est identifiée et que sa structure, bien que variable, est figée dans la définition de type de document (DTD, *Document Type Definition*).

Toutefois, hors du cartouche d'en-tête de chaque article encyclopédique, le texte devient libre et ce type d'approche montre alors sa limite. De plus, malgré une structure relativement rigide quant au balisage, l'information-même contenue dans les balises conserve une certaine latitude de présentation que le *wrapper* ne peut gérer du fait de sa grande rigidité, du fait aussi qu'il ne tient absolument pas compte du contenu informatif du texte qu'il a pour

charge d'extraire, mais qu'il s'arrête au repérage des bornes sans s'occuper de l'information elle-même [4]. Enfin, le fait que nous disposions de la DTD qui a servi à constituer l'encyclopédie rend redondante un travail d'induction sur sa structure déjà connue.

```
< Reference >
  < Entree.dyn >
    < Entree.ency >
      < Sommaire id = "e0000678.1" type = "T" >
        < Nom >< Format typo = "bold" >
          &Eacute; douard I&sup1e;
        < /Format >< /Nom >
      < /Sommaire >
      (...)
      < Etatcv > (
        < Lieu type = "naissance" >
          Westminster
        < /Lieu >
        (...)
        < Lieu type = "mort" >
          &#8212; Burgh by Sands, pr&egrave;s de Carlisle
        < /Lieu >
        (...)
      < /Etatcv >
      (...)
    < /Entree.ency >
  < /Entree.dyn >
  < Copyright value = "Hachette - Multimedia" / >
< /Reference >
```

FIG. 1.6 – *Encyclopédie Hachette Multimédia* : exemples de variations du texte structuré.

Dès lors, deux options se présentent à nous : soit utiliser une autre technique de gestion de l'information pour la partie la plus structurée de chaque article, soit ajouter au *wrapper* de [Kushmerick et al., 1997] une méthode

[4] L'exemple 1.6 montre que même dans sa partie structurée, le texte de l'*Encyclopédie Hachette Multimédia* subit des variations d'un article à l'autre. Si les éléments en italiques correspondent bien à l'information prévue par la balise (`Nom` et `Lieu` de `naissance`), le texte en caractères gras s'en éloigne en donnant des informations supplémentaires qu'il s'agit de traiter.

qui permettrait de pénétrer à l'intérieur de l'information, ou en tout cas un traitement postérieur à son extraction.

L'utilisation d'expressions régulières : *WHISK*

La seconde approche permettant d'effectuer une extraction d'information sur des textes structurés est *WHISK* [Soderland, 1999]. Basé sur la technologie des **expressions régulières** (*Regular Expressions*), cette méthode ne se limite pas aux textes formatés mais, plus souple que *Wrapper Induction*, elle traite également les documents semi-structurés. De plus, sous une forme étendue et plus évoluée, elle aborde également le texte libre. Nous nous limitons ici aux textes structurés et semi-structurés, pour lesquels la démarche reste la même. Nous aborderons le traitement du texte libre dans la section 1.2.4 page 46.

Comme pour *Wrapper Induction*, le principe de cette méthode est d'identifier, dans des textes pré-étiquetés, le contexte direct de l'information désignée comme pertinente, et en particulier les délimiteurs de cette information, pour constituer des règles d'extraction. Dans *WHISK*, ces règles se présentent sous forme d'expressions régulières constituées en patrons. Un patron correspond à une ou plusieurs expressions régulières, dont chaque mémorisation coïncide avec un élément de l'information pertinente que le patron doit fournir.

Les expressions régulières conviennent bien pour définir de tels patrons. En effet, elles sont capables de décrire rigoureusement les délimiteurs de l'information, mais elles présentent également l'avantage de pouvoir assouplir les contraintes sur ces délimiteurs, et même sur le texte compris entre les délimiteurs si cela est nécessaire. De plus, il est possible de définir plusieurs champs d'extraction dans une même expression régulière, et donc d'effectuer une extraction combinée.

Les règles définies par [Soderland, 1999] présentent toutefois quelques particularités si on les compare aux expressions régulières classiques définies par [Aho et Ullman, 1973, Aho et al., 1988]. Essentiellement, les quantificateurs (« * » et « + ») sont non gourmands, c'est-à-dire que leur étendue de fonctionnement n'est pas la plus longue possible. Cela permet de favoriser la proximité entre les éléments d'une même information dans le cadre d'une extraction combinée, et de limiter le temps de traitement du document en évitant de tenter des mises en correspondance très longues.

D'autre part, la capacité des expressions régulières à manipuler des classes de caractères (par exemple [a-z], [0-9]...) a été étendue de manière à ce que des classes de mots puissent être utilisées. Il est donc possible de définir des ensemble de mots jugés équivalents et de les regrouper sous une seule ap-

Exemple de texte semi-structuré : annonce en ligne proposant un appartement en location [a] :

```
Capitol Hill - 1 br twnhme. fplc D/W W/D.
Undrgrnd pkg incl $675. 3BR, upper flr
of turn of ctry HOME. incl gar, grt N. Hill
loc $995 (206) 999-999 <br>
<i><font size=-2> (This ad last ran on 09/03/97.)
</font></i><hr>
```

Exemple de règle d'extraction *WHISK* pour obtenir le nombre de chambres et le prix :

```
ID : : 1
Pattern : : *(Digit) ' BR' * '$' (Number)
Output : : Rental {Bedrooms $1} {Price $2}
```

Réponse produite par la règle *WHISK* sur le texte ci-dessus :

```
Rental :
       Bedrooms :              1
       Price :                 675

Rental :
       Bedrooms :              3
       Price :                 995
```

[a] Les éléments d'information extraite par la règle d'extraction *WHISK* sont en italique. Les endroits de correspondance du patron sont en gras.

FIG. 1.7 – Forme et fonctionnement d'une règle *WHISK* sur du texte semi-structuré.

pellation, un mot-clef qui apparaît en italiques dans le patron d'extraction. Dans l'exemple 1.7, le patron d'extraction comporte un *Digit* qui représente un chiffre, et un *Number* qui désigne un nombre d'un chiffre ou plus.

Les règles d'extraction de *WHISK* fonctionnent comme suit : lorsque chaque élément mémorisé par le patron correspond à une portion du texte, ces éléments sont mémorisés ; si le texte n'est pas entièrement parcouru par

le patron qui s'est appliqué, la même règle d'extraction est appliquée à la portion du texte restante (dans l'exemple, un même patron est appliqué deux fois) ; lorsqu'il y a échec d'un patron, mais que des éléments de la règle ont pu être identifiés, on conserve ces éléments et on relance la même règle sur le texte au-delà des éléments identifiés pour éviter de mélanger des éléments d'information distincts.

Les particularités des règles de *WHISK* permettent de gérer dans une certaine mesure la sémantique dont peut être porteur le texte structuré ou semi-structuré, et donc de faire face aux variations qui lui sont propres. En effet, la possibilité de manipuler des classes de mots permet de se détacher de la contrainte lexicale pour se concentrer sur des éléments linguistiques de plus haut niveau. Cela permet également de généraliser les règles bien plus aisément que ne l'autorise *Wrapper Induction*. D'autre part, l'utilisation des expressions régulières dans les patrons d'extraction permet d'appliquer des contraintes plus variées que les règles HLRT de *Wrapper Induction*, et notamment de tenir compte du contenu-même des groupes de mémorisation, comme c'est le cas pour le nombre des chambres et le prix de location dans l'exemple. Les expressions régulières sont donc bien plus adaptées que ces règles à l'extraction d'une information certes balisée, mais irrégulière à l'intérieur des balises. Cette constatation est intéressante car les outils dont nous disposons permettent de manipuler des expressions régulières. Il ne s'agit toutefois que de reconnaître une information pertinente dans un texte à partir de patrons, et pas encore d'identifier un sens ou une signification dans ce texte.

L'approche multistratégie de [Freitag, 1998]

Comme *WHISK*, l'approche de [Freitag, 1998] permet de traiter des documents qui ne suivent pas les règles grammaticales d'une langue, qu'ils soient structurés ou semi-structurés. Elle repose sur la mise en œuvre de trois stratégies d'extraction de l'information et se concentrent sur un seul champ d'extraction. Il s'agit donc d'une extraction simple. Le principe de fonctionnement de cette approche est le suivant : tous les fragments[5] du texte sont considérés comme des réponses possibles depuis les plus petits (avec comme limite le nombre de mots de la réponse la plus courte dans le corpus d'entraînement) jusqu'au plus grand (avec comme limite le nombre de mots de la réponse la plus longue). Un indice de confiance est appliqué à chacun des groupes selon les trois méthodes suivantes.

[5] Un fragment de texte n'est qu'une séquence d'unités lexicales qui se suivent dans le texte, sans autre contrainte relationnelle.

La plus simple des méthodes d'extraction exploitée consiste à mémoriser par une procédure rudimentaire d'apprentissage par cœur (*rote learner*), dans le corpus d'entraînement, le fragment de texte constituant la réponse à la requête prédéfinie afin de construire un dictionnaire d'exemples de réponses. Pour la perfectionner et donner à chaque fragment une pondération de confiance, on estime la probabilité P pour chaque fragment f d'être une réponse pertinente à la requête (R_f est le nombre d'apparition du fragment de texte dans une réponse ; T_f est le nombre d'apparition total du fragment de texte) :

$$P(f) = \frac{R_f+1}{T_f+2}$$

L'ajout de 1 au dividende et de 2 au diviseur s'explique par le fait qu'il ne faut pas exclure la possibilité statistique qu'un fragment de texte soit une réponse pertinente à une requête proposée simplement parce que ce fragment n'apparaît pas dans les réponses du corpus d'entraînement ou même dans le corpus d'entraînement lui-même (formule de Laplace). Ainsi, *P(f)* correspondra à une fraction de 1 si f n'est pas une réponse pertinente dans le corpus d'entraînement, et *P(f)* aura une valeur de 0.5 si ce même énoncé f n'apparaît pas dans le corpus, ce qui équivaut à ne pas prendre de décision.

La deuxième méthode, héritée de la classification de documents, est une généralisation de la stratégie précédente. Mais, basée sur l'utilisation de la formule de la probabilité des causes de Bayes (*Bayes learner*), elle permet de tenir compte de son environnement lexical puisqu'elle attribue à chaque mot une probabilité d'appartenir à une réponse en fonction de son contexte, chaque réponse du corpus d'entraînement étant considéré comme un sac de mots (*bag-of-words*). Chaque fragment d'un texte est considéré comme une hypothèse (H) à vérifier et son environnement (D) influe sur sa pertinence :

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_{j=1}^{n} P(D|H_j)P(H_j)}$$

Ces deux approches présentent peu d'intérêt pour le problème qui nous occupe, puisqu'elles se bornent à exploiter des données statistiques et des mesures de fréquences. La troisième méthode présentée par [Freitag, 1998] (*SRV*, *relational learner*) tient compte quant à elle d'indices linguistiques (syntaxe, morphologie) en plus d'autres caractéristiques (présentation du document). Ces traits sont appelés simples s'ils s'appliquent à un seul mot (par exemple `capitalized?` ou `noun?`) et relationnels s'ils concernent le comportement d'un mot vis-à-vis d'un autre (par exemple `next-token` ou `subject-verb`). Il est dès lors possible d'induire des règles relationnelles dont ils constituent les contraintes.

L'exemple 1.8 page ci-contre présente une règle qui recherche un ou des éléments `A` précédés par un mot en capitales deux mots plus tôt dans le texte.

```
some( ?A [prev-token prev-token]
capitalized true)
```

FIG. 1.8 – Exemple de règle relationnelle SRV.

La méthode du système la plus proche de nos préoccupations linguistiques est *SRV*, qui génère des règles de contraintes pour effectuer l'extraction. Elle est intéressante dans son approche de traits à la fois de nature linguistiques et autres, comme la présentation typographique, mais reste limitée à une extraction simple et nous ne voyons pas comment la perfectionner pour l'ouvrir à une extraction combinée. D'autre part, les règles relationnelles ne présentent pas une grande originalité par rapport aux patrons que l'on a observés dans le système *WHISK*. En effet, les expressions régulières sont à même de prendre en compte les mêmes indices tout en conservant une extraction combinée. De plus, elles constituent depuis longtemps un standard dans le monde de l'informatique, et plus particulièrement dans celui du traitement automatique des langues. Il ne s'agirait donc pas d'implanter un formalisme neuf, ce qui constitue un grand avantage pratique. Toutefois, [Freitag, 1998] suggère à bon escient une approche multistratégie du problème car il est difficile de traiter de même façon du texte grammatical et de l'information structurée.

L'exploitation de syntaxe et de sémantique par RAPIER

[Califf et Mooney, 1997, Califf et Mooney, 1999] décrivent RAPIER (*Robust Automated Production of IE Rules*), un système de génération de règles d'extraction destiné, dans son état actuel, à du texte qui ne respecte pas la grammaire d'une langue. Les règles d'extraction de RAPIER utilisent des patrons qui exploitent des indices syntaxiques limités (partie du discours) et sémantiques (lexique sémantique tel que *WordNet* [Fellbaum, 1998b]). On repère dans chacun des textes du corpus d'entraînement l'information qui correspond à un champ d'extraction et de ce fait on divise le texte en trois partie :

1. la partie du texte qui précède la réponse ;
2. la réponse à une requête prédéfinie ;
3. la partie du texte qui suit la réponse.

Les règles de RAPIER sont des patrons d'extraction correspondant à ces différents champs. D'un point de vue lexical, le patron correspond à une liste plus ou moins longue de **lexèmes** appartenant à chacun des champs (on peut faire varier ce nombre). Les indices syntaxiques et sémantiques sont ajoutés au patron selon qu'ils correspondent ou non aux listes extraites pour chacun des champs. Le type de règles mises en œuvre ne permet pas d'effectuer une

extraction combinée (puisqu'il n'y a qu'un champ informatif) et il semble que le perfectionnement de cette approche demande une importante complication de la méthode.

Pre-filler :	Filler :	Post-filler :
1) tag : nn,nnp	1) word : undisclosed	1) sem : price
2) list : length 2	tag : jj	

TAB. 1.1 – Exemple de règle d'extraction de RAPIER.

L'exemple 1.1 montre une règle construite par RAPIER pour extraire le nombre de transactions indiqué dans des communiqués concernant des acquisitions d'entreprises. Cette règle extrait la valeur `undisclosed` de phrases telles que *sold to the bank for an undisclosed amount* ou *paid Honeywell an undisclosed price*. Le patron qui précède la réponse contient deux éléments : un mot dont la **catégorie grammaticale** est soit un nom (`nn`) soit un nom propre (`nnp`), et une liste de deux mots au moins sans contrainte. Le patron d'extraction contient le mot `undisclosed` étiqueté comme adjectif (`jj`). Le patron suivant la réponse requiert un mot dont la catégorie sémantique (dans *WordNet*) est `price`.

Le format des règles d'extraction de RAPIER impose un nouveau patron pour chaque type d'information à identifier, ce qui n'autorise pas d'extraction de données combinées. D'autre part, bien que la procédure reste sensiblement la même que dans les autres méthodes observées (contraintes diverses sur les bornes gauche et droite de l'information, ainsi que sur le champ informatif lui-même), le formalisme de la règle est encore une fois spécifique à l'application et ne présente pas la même souplesse que les expressions régulières standard, sans pour autant offrir de fonctionnalité particulière. Comme ses homologues, RAPIER demande une phase d'apprentissage sur un corpus d'entraînement étiqueté, et se limite à des requêtes prédéfinies.

1.2.4 Méthodes d'extraction pour les textes libres

Les systèmes d'apprentissage de règles d'extraction pour le texte libre sont assez rares et se cantonnent souvent à une désambiguïsation sémantique lexicale locale au mot, ou à une étude statistique du lexique des documents [Sheridan et Ballerini, 1996, Gaussier et al., 2000]. D'autres approches nous intéressent, plus proches de l'étude linguistique du texte. Ce sont ces approches que nous présentons ici.

AutoSlog (et *AutoSlog-TS*) avec *CIRCUS*

Une des premières approches à dominante linguistique en extraction d'information s'est effectuée au travers de l'analyseur conceptuel *CIRCUS* [Lehnert, 1990]. Il s'agit d'un outil d'analyse de texte qui se base à la fois sur une analyse syntaxique sommaire (ni arbre, ni grammaire) et sur un dictionnaire de nœuds de concept (*concept nodes*) pour produire sous forme d'un tableau d'information (*case frame*) une représentation sémantique d'un texte présenté en entrée.

Chaque entrée de ce dictionnaire de nœuds de concept décrit un événement, c'est-à-dire qu'il définit les contraintes syntaxiques et sémantiques (lexicales) qui doivent être remplies par le texte pour correspondre à cette entrée. On dit qu'une entrée est activée lorsque le texte remplit les contraintes qu'elle fixe. Un tableau informationnel est la structure de l'information du texte telle qu'elle se présente lorsque un ou des nœuds conceptuels ont été activés.

Dans l'exemple 1.9 page suivante, la phrase déclenche trois nœuds conceptuels du dictionnaire : la construction passive du verbe *to destroy* permet d'en identifier le sujet comme cible de destruction ; la construction passive du verbe *to damage* permet également d'en sélectionner le sujet comme cible ; enfin, le mot *bomb* est lexicalement identifié comme une arme. À partir de ces nœuds déclenchés, le *case frame* correspondant peut recevoir diverses informations : les signatures (constructions relatives aux déclencheurs) ainsi que les cibles et l'instrument.

Il faut cependant noter que *CIRCUS* réclame un dictionnaire de nœuds de concept spécialisé pour chaque domaine d'extraction. [Riloff, 1993] note que ce dictionnaire de nœuds de concept est extrêmement long à réaliser (environ 1 500 heures/personne) et qu'il est utopique de croire pouvoir en obtenir un pour chaque domaine dans lequel on voudrait réaliser de l'extraction d'information, ce qui réduit la portabilité de *CIRCUS*. Aussi Ellen Riloff propose-t-elle le système *AutoSlog* [Riloff, 1993, Riloff, 1996b, Riloff et Shepherd, 1997] capable, à partir d'un corpus représentatif d'un domaine, de construire un dictionnaire de nœuds de concept spécialisé de ce domaine. Cette approche repose sur deux observations :

1. les informations les plus pertinentes dans le cadre d'un *case frame* apparaissent lors de la première mention de l'événement qu'elles caractérisent [6] ;
2. chacune de ces informations comporte dans son contexte linguistique des indices qui en décrivent le rôle dans l'événement concerné.

Il ressort de ces observations que la mise en exergue d'une **dépendance**

[6] Cette observation concerne les corpus de MUC-3. Il est à craindre qu'elle ne soit pas générale, mais qu'elle s'applique plus particulièrement à ces corpus.

```
 Sentence :
Two vehicles were destroyed and an unidentified office of
the agriculture and livestock ministry was heavily damaged
following the explosion of two bombs yesterday afternoon.

Concept nodes
      $destruction-passive$ (déclenché par destroyed)
            target = « two vehicles »

      $damage-passive$ (déclenché par damaged)
            target = « an unidentified office of the agriculture and
            livestock ministry »

      $weapon-bomb$ (déclenché par bombs)

Case frame
      Signatures :
            (<destroyed, $destruction-passive$>, <damaged,
            $damage-passive$>, <bombs, $weapon-bomb$>)
      Perpetrators :
            nil
      Victims :
            nil
      Targets :
            (GOVT-OFFICE-OR-RESIDENCE TRANSPORT-VEHICLE)
      Instruments :
            (BOMB)
```

FIG. 1.9 – Exemples de texte, de nœuds de concept et d'un tableau informationnel.

syntaxique entre deux termes laisse supposer un rapport sémantique entre ces mêmes termes.

Pour concevoir un dictionnaire des nœuds de concept représentatif d'un domaine, le corpus d'apprentissage doit identifier les éléments de l'information appropriée pour remplir un *case frame* donné correspondant à l'information pertinente du domaine. Pour chaque texte de ce corpus, les éléments d'information en sont listés et identifiés. Ainsi, si la phrase de l'exemple (*cf.* figure 1.9) appartient à un document d'un tel corpus, certains éléments en seront identifiés et typés par un spécialiste comme étant pertinents pour la recherche d'information pour le domaine des attaques terroristes : *two vehicles* et *an office of the agriculture and livestock ministry* sont des cibles (*targets*), tandis que *bombs* est un instrument.

À l'aide du corpus d'entraînement ainsi traité, *AutoSlog* va générer les nœuds de concept selon la démarche qui suit :

- *AutoSlog* identifie la première phrase contenant le syntagme à identifier (par exemple la première cible *two vehicles*) ;
- *CIRCUS* effectue une analyse conceptuelle de la phrase ;
- *AutoSlog* identifie dans le résultat de l'analyse conceptuelle le groupe qui contient le syntagme et les relations dont il dépend ;
- *AutoSlog* vérifie la correspondance entre la structure syntaxique à laquelle appartient le syntagme et une au moins des heuristiques (*cf.* table 1.2) qui lui sont propres. Une heuristique correspond à une structure syntaxique porteuse d'information : <*subj*> *passive-verb* détermine une information sur le sujet d'un verbe dans une construction passive ;
- lorsqu'aucune des heuristiques n'est satisfaite, *AutoSlog* cherche la phrase suivante dans laquelle apparaît le syntagme et reprend la démarche au début. Si une heuristique (au moins) est satisfaite, *AutoSlog* présente à un opérateur expert la ou les propositions de définition pour le dictionnaire.

Heuristique	Exemple
`<subject> passive-verb`	<victim> was murdered
`<subject> active-verb`	<perpetrator> bombed
`<subject> verb infinitive`	<perpetrator> attempted to kill
`<subject> auxiliary noun`	<victim> was victim
`passive-verb <dobj>`	killed <victim>
`active-verb <dobj>`	bombed <target>
`infinitive <dobj>`	to kill <victim>
`verb infinitive <dobj>`	threatened to attack <victim>
`gerund <dobj>`	killing <victim>
`noun auxiliary <dobj>`	fatality was <victim>
`noun prep <np>`	bomb against <target>
`active-verb prep <np>`	killed with <instrument>
`passive-verb <np>`	was aimed at <target>

TAB. 1.2 – Les heuristiques d'*AutoSlog*.

Avec *AutoSlog*, nous avons la présentation d'un système de génération de règles d'extraction basé sur les relations syntaxico-sémantiques entre les mots tout en privilégiant les données lexicales à travers l'extraction de données d'un corpus d'entraînement. De la sorte, l'utilisation d'une ressource lexicale indiquant la sémantique des mots et leur construction syntaxique se révèle attrayante et intéressante. [Soderland, 1999] émet des réserve sur le niveau de précision de la réponse fournie par *AutoSlog*, qui se contente de restituer des groupes syntaxiques complets. Cette réticence n'a pas de raison d'être dans

notre approche, qui cherche à fournir une réponse au niveau de la phrase, du paragraphe voire du document si les éléments informatifs recherchés sont disséminés dans plusieurs phrases.

Cependant, la limitation du champ d'application d'*AutoSlog* à une seule information par document du fait d'un mode d'extraction simple en réduit grandement la portée. De plus, même si la création du dictionnaire lors de la mise en œuvre d'*AutoSlog* est infiniment plus rapide que la construction manuelle d'une telle ressource [7], il demeure qu'un travail important doit être mené sur le corpus d'entraînement lui-même. Ce travail n'entre pas dans la comptabilisation de [Riloff, 1993], alors qu'il relève d'un spécialiste et ne peut être mené très rapidement.

Cette remarque n'a pas échappé à l'attention des concepteurs d'*AutoSlog* : en effet, [Riloff, 1996a] note que 50 heures/personne environ sont nécessaires pour annoter 1 000 documents. [Riloff, 1996a, Riloff et Lorenzen, 1999] définissent un nouveau système, *AutoSlog-TS*, qui peut se contenter d'exploiter un corpus représentatif d'un domaine, avec la seule indication de présence ou non d'une information pertinente dans chaque document afin d'éviter les problèmes liés à la conception d'un corpus étiqueté. Cependant le système devient alors un simple catégoriseur de documents, et il sort dès lors de notre champ d'intérêt puisqu'il n'est plus capable d'extraire une information précise.

La généralisation des contraintes dans *CRYSTAL*

Avec *CRYSTAL*, [Soderland et al., 1995] s'inscrivent bien dans la continuité des travaux menés par [Riloff, 1993] pour *AutoSlog*. En effet, *CRYSTAL* repose également sur l'utilisation d'un système d'extraction d'information qui s'appuie sur un dictionnaire de nœuds conceptuels. Ce système d'extraction est un analyseur de phrases appelé *BADGER*, qui a succédé a *CIRCUS* a l'Université du Massachusetts. De la même manière que le système *AutoSlog* aussi, *CRYSTAL* repose sur le principe de la construction automatisée d'un dictionnaire de nœuds conceptuels à partir d'un ensemble manuellement préétiqueté d'exemples d'informations pertinentes dans le domaine visé.

Ces nœuds conceptuels ont toutefois évolué depuis ceux de *CIRCUS* : moins soumis à des types d'information prédéfinis, ils s'appuient sur une caractérisation de l'information locale au nœud, et plus à une caractérisation attachée au dictionnaire dont il fait partie. D'autre part, l'extraction combinée est maintenant possible. L'exemple (*cf.* figure 1.10 page suivante) indique les différentes fonctionnalités qui leur ont été ajoutées.

[7] Un dictionnaire construit à la main demande 1 500 heures/personne là où *AutoSlog* ne réclame que 5 heures/personne.

```
 Sentence :
Unmarkable with the exception of mild shortness of breath
and chronically swollen ankles.

Concept node definition
       CN-type : Sign or Symptom
       Subtype : Present
       Extract from Prep. Phrase ''WITH''
       Verb = <NULL>
       Subject Constraints :
              words include ''UNREMARKABLE''
       Prep. Phrase Constraints :
              preposition = ''WITH''
              words include
                     ''THE EXCEPTION OF MILD SHORTNESS OF
                     BREATH AND CHRONICALLY SWOLLEN ANKLES''
              modifier class <Sign or Symptom>
              head class
                     <Sign or Symptom>,<Body Location or Region>
```

FIG. 1.10 – Definition de noeud de concept initial utilisant les unités lexicales de l'exemple.

D'autre part, le système *BADGER* permet d'exploiter plus profondément la syntaxe puisque les possibilités d'analyse ont été grandement améliorées depuis le précédent système.

Mais la principale amélioration de *CRYSTAL* repose dans sa faculté à généraliser les nœuds de concept extrêmement contraints qu'il extrait dans un premier temps. En effet, l'algorithme qui le pilote permet d'étendre ces contraintes tant sémantiques que syntaxiques au maximum, la limite étant la possibilité d'exploiter toujours avec la même efficacité le corpus d'entraînement qui lui est utile pour créer son propre dictionnaire de nœuds de concept. L'algorithme que nous reprenons ici (*cf.* figure 1.11 page suivante) indique de quelle manière le système procède.

Ce système, bien que présentant les avantages de la généralisation automatique et de l'extraction combinée par rapport au précédent système, n'apporte pas d'avancée significative dans le domaine de l'exploitation de l'information. Tout au plus valide-t-il l'utilisation d'une étude syntaxique plus évoluée et plus approfondie.

Initialiser le dictionnaire
Initialiser la base d'exemples d'entraînement
Aussi longtemps que pas plus de nœud de concept initial dans le dictionnaire
  D = une définition de nœud conceptuel initial retirée du dictionnaire
  Boucle
    D' = la définition de nœud de concept la plus proche de D
    Si D' = D, sortie de boucle
    U = unification de D et D'
    Test de couverture de U dans les exemples d'entraînement
    Si le taux d'erreur de U > tolérance, sortie de boucle
    Effacer toutes les définitions de noeuds de concept couverts par U
    D = U
  Ajouter D au dictionnaire
Retourner le dictionnaire

FIG. 1.11 – Algorithme d'extraction d'un dictionnaire de définitions de nœuds de concept par *CRYSTAL*.

L'utilisation de la syntaxe par *WHISK* pour le texte libre

WHISK est le seul système conçu pour traiter à la fois les documents structurés ou semi-structurés et le texte libre. Si son approche du texte libre subit des changements par rapport aux autres types de textes, le principe de fonctionnement de l'application reste le même.

En effet, c'est toujours une étude du contexte de l'information étiquetée qui permet la constitution de patrons sous forme d'expressions régulières. Toutefois, à l'image d'*AutoSlog* et de *CRYSTAL*, *WHISK* intègre une analyse syntaxique des énoncés qui permet d'identifier les rapports syntaxique de l'information avec son contexte et d'ajouter ce critère aux contraintes du patron.

La phase d'apprentissage au cours de laquelle sont construites les règles n'est toutefois pas complètement automatique. En effet, les textes (étiquetés) ne sont pas présentés tels quels au système. Un prétraitement est généralement nécessaire, notamment pour effectuer une **segmentation**. Les textes présentés à *WHISK* en sont des extraits correspondant à un exemple d'information par extrait.

La figure 1.12 page suivante présente un exemple de formation de règle d'extraction à partir d'un exemple en texte libre. L'analyse syntaxique de cet exemple permet de distinguer les champs syntaxiques (`SUBJ` pour sujet,

Énoncé :
C. Vincent Protho, chairman and chief executive officer of this maker of semiconductors, was named to the additional post of president, succeeding John W. Smith, who resigned to pursue other interests.

Analyse syntaxique :

```
@S[
     {SUBJ     @PN[C. Vincent Protho]PN , @PS[chairman
               and chief executive officer]PS of this
               maker of semiconductors,}
     {VB       @Passive was named @nam}
     {PP       to the additional post of
               @PS[president]PS , }
     {REL_V    succeeding @succeed @PN[John W. Smith]PN
               , who resigned @resign to pursue @pursu
               other interests.}
]@S 8910130051-1
```

Règle correspondante :

```
ID : : 3
Pattern : :  * ( Person ) * '@Passive' *F 'named' * {PP
             *F ( Position ) * '@succeed ' ( Person )
Output : :   Succession {PersonIn $1}{Post $2}{PersonOut
             $3}
```

FIG. 1.12 – Formation d'une règle *WHISK* pour du texte libre.

`VB` pour verbe principal, `PP` pour groupe prépositionnel et `REL_V` pour proposition relative rattachée à un verbe). Les indications précédées par `@` sont d'ordre sémantique (`PN` pour nom de personne, `PS` pour poste, `CN` pour nom de société), syntaxiques (`Passive` pour passif) ou morphologiques (la racine d'un verbe : `nam`, `succeed`, `resign`, `pursu`). La règle qui en découle est une expression régulière. Les mots en italiques indiquent des classes sémantiques. Les `@` permettent de requérir une indication morphologique (`@succeed` demande un verbe de radical *succeed*) ou syntaxique (`@Passive` exige un verbe à la voix passive). Le quantifieur `*` permet de ne pas prendre en considération un nombre illimité de caractères. S'il est suivi de `F` (`*F`), la fin du champ syntaxique sert de limite aux caractères dont il ne faut pas tenir compte.

L'application de telles règles requiert une analyse syntaxique des textes à traiter. Toutefois, les exigences syntaxiques présentes dans les règles peuvent être désactivées, et dès lors la règle fonctionnera comme pour du texte semi-structuré. Notons encore que les classes sémantiques ne sont pas toujours suffisantes pour identifier certaines entités. Des listes d'entités nommées et

des stratégies de reconnaissance de ces entités peuvent être mises en œuvre pour les reconnaître.

1.2.5 La campagne d'évaluation TREC

Suite au succès rencontré par MUC, diverses autres conférences ont vu le jour, qui abordent des disciplines semblables ou connexes à l'extraction d'information. En particulier, la campagne d'évaluation TREC (*Text REtrieval Conference*) a obtenu un consensus dans le domaine de la recherche d'information. TREC fait actuellement autorité pour tester la valeur des approches qui visent à la sélection de documents qui contiennent une information déterminée. Les systèmes qui concourent cherchent en effet à déterminer dans une base documentaire les documents qui correspondent à une information réclamée par un utilisateur [Harman, 1992].

Au cours des éditions successives de TREC[8], différentes tâches ont été définies qui répondent à des besoins réels d'application réclamées par le public. Ces tâches correspondent à différentes facettes de la gestion de l'information.

- Recherche multilingue.
- Filtrage.
- Recherche interactive.
- Analyse de requête.
- Question-réponse.
- Recherche en documents oraux.
- Recherche sur le Web.

Les mesures d'évaluation des systèmes présentés dans les campagnes d'évaluation TREC correspondent aux **mesures traditionnelles** de rappel et de précision, qui sont par ailleurs utilisées dans le cadre des conférences MUC (*cf.* section 1.2.1 page 26). Seule la tâche de question-réponse, très particulière, fait exception car le rappel n'a pas été jugé prépondérant pour ce type d'application. C'est donc un **score** correspondant au rang de la première bonne réponse pour chaque question qui indique la qualité du système. Par ailleurs, les lacunes du système sont indiquées par le nombre de questions qui n'obtiennent pas de bonne réponse (*cf.* section 7.2.2 page 201).

Parmi les différentes perspectives offertes par TREC, c'est la tâche de question-réponse qui a particulièrement retenu notre attention. En effet, contrairement aux autres, elle exige une **fenêtre** de réponse inférieure au document entier et réclame une identification plus ou moins exacte de l'information recherchée[9]. De plus, les questions ne sont pas limitées à une in-

[8] TREC est une campagne d'évaluation annuelle. En novembre 2002 a eu lieu la onzième édition.

[9] Les huitième et neuvième éditions de TREC prévoyaient deux fenêtres de réponse de 50 et 250 caractères. La dixième édition a supprimé la fenêtre de 250 caractères. La dernière édition supprime l'arbitraire d'une fenêtre limitée pour demander seulement la

formation ou à un type d'information. Dès lors, la tâche de question-réponse affirme résolument son caractère généraliste. Ces deux caractéristiques particulièrement exigeantes nous amènent à nous intéresser à la tâche de question-réponse comme nous l'avons fait pour celle d'extraction d'information.

Notre propos n'est pas ici d'étudier les méthodes de question-réponse existantes, mais plus élémentairement les techniques qui permettent d'identifier une information et de la traiter, avant toute localisation d'informations correspondantes dans les textes et dans les requêtes. L'objectif de notre thèse est en effet d'élaborer une méthodologique de construction d'une structure informationnelle à partir d'une base documentaire. Cette structure informationnelle doit permettre de gérer l'information contenue dans la base documentaire quels que soient les besoins de l'utilisateur. L'interrogation particulière de la base documentaire constitue une évaluation de la qualité de la structure, mais elle reste partielle et dirigée. Le but est en effet d'obtenir une méthodologie généraliste reposant sur des méthodes linguistiques.

Dès les années septante, la problématique de question-réponse a été envisagée et traitée grâce à des approches de type linguitique. À cette époque, le genre des textes et le domaine auquels ils appartenaient étaient extrèmement spécifiques. Par exemple, le système *QALM* [Lehnert, 1977, Lehnert, 1979] analyse de courtes histoires sur des sujets très précis et limités pour en extraire une représentation conceptuelle. Le système *QALM* dispose en outre d'une base de connaissances propres au domaine du scénario analysé, ainsi que d'une typologie des questions disposant de 13 catégories de question qui possèdent leur propre heuristique pour trouver la réponse à la question proposée. Ces heuristiques reposent sur une analyse du contenu de la question, sur une recherche dans la représentation conceptuelle du scénario et sur un raisonnement à partir de la base de connaissances.

Les systèmes ultérieurs ne différent de *QALM* que par une extension des connaissances, notamment pragmatiques, de l'univers appréhendé, et par une plus grande variété de types de questions [Dyer, 1983, Zock et Mitkov, 1991]. Le système *QUEST* [Graesser et al., 1994], qui correspond à la même approche, définit les quatre composantes de ce type d'architecture :

- catégorisation des questions ;
- identification des sources d'information qui permettent de répondre (liées au domaine ou génériques) ;
- processus de mise en correspondance des faits et événements des questions et de propositions de réponse ;
- formulation de la réponse.

[Ferret et al., 2002a] estime que cette architecture exclusivement linguistique n'est pas réalisable pour une application généraliste car les sources d'information devraient alors comprendre une définition et une formalisation des

réponse à la question posée.

connaissances pragmatiques sans limite de domaine. [Mollá Aliod et al., 2000] ne dit pas autre chose lorsqu'il utilise un modèle semblable pour poser des questions sur les commandes UNIX, tout en adjoignant à une analyse syntaxico-sémantique un raisonnement logique reposant sur des inférences liées à un lexique limité par le domaine et à des connaissances sémantiques du domaine.

Et en effet, depuis la première édition de TREC, la plupart des systèmes de question-réponse généralistes sont basés sur une architecture légèrement différente :

- catégorisation des questions ;
- moteur de recherche permettant une première sélection des documents ou fragments de documents susceptibles de répondre à la question posée ;
- traitements linguistiques et autres appliqués sur les documents sélectionnés pour déterminer les réponses possibles.

Les différences reposent essentiellement dans les traitements d'analyse de la question et dans ceux des textes. Les traitements appliqués à la question sont propres aux méthodologies de question-réponse. Les procédés appliqués aux textes présélectionnés correspondent à un traitement de l'information contenue dans ces textes. Le moteur de recherche permettant de sélectionner les textes candidats n'appartient pas à la méthodologie de question-réponse, mais les processus utilisés pour sélectionner des documents qui ne contiennent pas forcément les unités lexicales contenues dans l'information extraite de la question ne doivent pas être négligés.

Dès la première évaluation des systèmes de question-réponse dans TREC [Voorhees, 1999], cette architecture a été mise en œuvre. Par exemple, le système de [Hull, 1999] analyse les questions pour en extraire le vocabulaire et pour en catégoriser l'objet grâce à l'interrogatif et à certains patrons lexicaux [10]. Le vocabulaire ainsi extrait permet de constituer un ensemble de textes qui lui correspondent grâce au système d'extraction d'information de AT&T.

Les textes extraits sont analysés et chacune des phrases de ces textes sont classifiées en fonction du nombre de mots qu'elles contiennent en commun avec la question. Les noms propres et les nombres reçoivent le poids le plus important, puis les noms communs ou inconnus. Les autres mots sont peu considérés. L'application d'un module de reconnaissance d'entités (*ThingFinder* [Trouilleux, 1998]) permet ensuite d'identifier les noms de personne, de lieu, les expressions de date, de prix, de quantité ou de nombre. Ces entités sont mises en correspondance avec le type de la question et les phrases qui ne

[10] Une question dont l'interrogatif est *How* sera normalement catégorisée `<How>`. Toutefois, si la catégorisation produit `<How>` *Adj* (où *Adj* est un adjectif), la nature de l'adjectif permet d'aboutir à d'autres catégories : si *Adj* est *long* ou *short*, les type sera `<Quantitiy>` ; si *Adj* est *rich* ou *poor*, le type sera `<Money>`.

contiennent pas le type attendu sont éliminées. Les mots ou expressions qui correspondent au type de la question sont considérés comme des réponses potentielles. Le vocabulaire contenu dans la question est éliminé.

Les résultats obtenus par ce système sont relativement honorables. Toutefois, l'auteur regrette à plusieurs reprises les erreurs que le manque de traitements linguistiques ne permet pas de corriger. Notamment, lors de son analyse des phrases sélectionnées, il déplore le manque de liens entre les réponses possibles et le contenu de la question. Par ailleurs, l'analyse de la question elle-même demande des ressources sémantiques dont le système ne dispose pas.

L'évolution des méthodes de question-réponse depuis la huitième édition de TREC en 1999 n'a pas modifié l'architecture générale des systèmes, qui sont toujours basés sur une catégorisation des requêtes, sur une recherche par mots-clefs dans les documents à l'aide d'un moteur de recherche généralement externe et sur des traitements des documents sélectionnés par le moteur pour en identifier ou en extraire la meilleure réponse. Le système *QALC* de [Ferret et al., 1999, Ferret et al., 2002b] s'appuie sur la constatation que les méthodes qui comportent les traitements linguistiques les plus élaborés sont également ceux qui atteignent les meilleures performances. Dès lors, et pour chaque partie du système, les traitements linguistiques sont privilégiés dans cette approche.

Tout d'abord, l'analyse de la question doit permettre d'obtenir deux informations. D'une part c'est grâce à elle qu'est atteinte la catégorisation de l'objet de la question, et donc de la réponse attendue. Cette catégorisation est réalisée par des patrons qui s'appuient sur des critères lexicaux (principalement la nature de l'interrogatif), syntaxiques (la catégorie syntaxiques des groupes en relation syntaxique directe avec l'interrogatif) et sémantiques (des catégories sémantiques fournies par *WordNet*). L'application d'un patron de catégorisation identifie la catégorie de la réponse attendue à la question parmi quinze étiquettes qui correspondent aux entités nommées. D'autre part, l'analyse de l'énoncé de la question permet d'identifier les mots qui la constituent, et plus particulièrement des expressions syntaxiques complexes, appelées termes de recherche [11]. Ces termes et mots sont appelés à servir de mots-clefs lorsque le moteur effectue sa recherche.

Si le choix d'un moteur de recherche dans le cadre de cette application est basé sur la capacité de ce moteur à fournir une bonne réponse dans le plus grand nombre de cas par rapport à ses concurrents ainsi qu'à présenter le plus grand nombre de bonnes réponses, il repose également sur son aptitude a prendre en compte divers phénomènes linguistiques, et notamment la **synonymie** et des techniques de **racinisation** (*stemming*). L'ensemble de

[11] Les termes sont des expressions lexico-syntaxiques non-lexicalisées qui constituent une unité sémantique selon les critères définis par [Justeson et Katz, 1995].

l'information extraite de la requête est donc exploitée pour sélectionner des documents contenant les mêmes données et donc susceptibles de contenir la réponse.

Enfin, divers traitements sont appliqués aux documents proposés par le moteur de recherche afin de déterminer plus précisément la réponse à la question et pour classifier les propositions de réponse en fonction du degré de similitude de la proposition avec l'énoncé de la question. Le premier traitement est effectué par l'analyseur transformationnel *Fastr* [Jacquemin, 1999] qui permet d'envisager un grand nombre de variations morphologiques (les mots de même racine que l'unité originale) et sémantiques (les mots contenus dans un ensemble synonymique (***synset***) de *WordNet* 1.6 où apparaît l'unité originale) de la question. Notons ici qu'aucune désambiguïsation sémantique n'est appliquée et que tous les *synsets* sont considérés. À partir des familles morphologiques et sémantiques, des patrons sont constitués qui peuvent identifier l'expression originale de la question et ses variation présentes dans les textes. Il est dès lors possible d'affecter un poids à chaque document, qui est fonction inverse de son degré de variation par rapport à l'énoncé de la question. La présence de noms propres et celle des termes les plus longs sont deux facteurs qui augmentent le poids accordé à un document. Les vingt documents les plus pertinents sont classifiés et conservés.

Le deuxième traitement consiste à déceler les entités nommées (personnes, organisation, lieux, valeurs) au sein des documents de la sélection. Pour ce faire, *QALC* exploite divers dictionnaires d'entités nommées, des lexiques sémantiques dont il adapte l'information et des règles dédiées à chaque type d'entité, utilisées lorsque les lexiques sont lacunaires. Au niveau numérique, ces règles distinguent les nombres cardinaux et ordinaux, les expressions complexes « nombre-unité » (distances, valeurs monétaires,...), les expressions de temps et les autres nombres. Les organisations sont dénotées par la présence d'unités lexicales déterminées (*Administration*, *Association*,...) tandis que les noms de personnes correspondent à des patrons lexicaux (*Dr*, *President*,...) ou typographiques (majuscules,...). L'identification de ces entités correspond à la catégorisation des questions.

Enfin, l'appariement de la question avec la réponse se fait au niveau de la phrase, qui présente une réponse courte dans un contexte suffisant pour juger de sa pertinence. Chaque phrase de chaque document proposé reçoit un score d'appariement en fonction de trois critères : la présence de mots simple de la question dans la phrase, la présence de termes ou d'une de leurs variantes dans la phrase, la présence des entités nommées dans la phrase. Chaque type d'entité présente à la fois dans la question et dans la phrase reçoit un poids qui lui est propre et le poids de chaque phrase correspond à la combinaison des poids de chaque type, les mots simples valant deux fois les termes et les entités nommées. Toutefois, une proposition dans laquelle

aucune entité ne correspond à la catégorie de la question est éliminée. La réponse la plus pertinente est celle dont le poids est le plus élevé.

Le système qui obtient les meilleurs résultats dans les différentes évaluations TREC des systèmes de question-réponse est aussi celui qui utilise les procédures d'analyse linguistique les plus élaborées. Il s'agit du système *Falcon* [Moldovan et al., 2000, Harabagiu et al., 2000]. Comme les autres systèmes de question-réponse, cette application procède en trois étapes : catégorisation de la question, application d'un moteur de recherche sur les documents, analyse des réponses proposées pour déterminer un ordre de pertinence.

D'abord, un analyseur probabiliste est chargé de repérer les dépendances entre les mots de la question. Le résultat de cette analyse permet de reformuler la question sous la forme d'un graphe relationnel qui relie les têtes de groupes. Ces dépendances sont anonymes, ce qui ne permet pas de juger de leur importance. Ce graphe, ou formulaire sémantique – car les unités lexicales sont reliées à la **taxinomie** de *WordNet* – permet non seulement d'identifier le type de la question (la tête qui à la plus grande connexion syntaxique), mais aussi les mots-clefs qui sont utilisés par le moteur de recherche (les noms directement reliés au type ainsi que les adjectifs et les adverbes). Le type lui-même appartient à une des 27 catégories d'entités nommées, traduite dans un des 15 nœuds hiérarchiques supérieurs de *WordNet* pour la recherche. Aucune désambiguïsation sémantique n'est effectuée.

Trois types d'alternances sont prévus pour pallier les variations de la réponse par rapport à la question. L'alternance peut être morphologique (flexions et dérivations de mots-clefs), lexicale (utilisation de synonymes) ou sémantique (termes semblables sans être synonymes, **hypéronymes**).

À partir des éléments extraits de la question, une recherche est lancée grâce à un moteur de recherche **booléen** qui permet les alternances proposées. La recherche est menée par paragraphe dans les documents. Les propositions du moteur de recherche sont en effet des paragraphes qui contiennent les entités les plus représentatives de la question et une entité correspondant au type de la question.

Enfin, les propositions du moteur de recherche sont soumises à l'analyseur probabiliste et un formulaire sémantique est construit. L'unification du formulaire de la question avec celui de la réponse est tenté, d'abord au niveau lexical, puis avec les alternances possibles. Lorsqu'une unification des formulaires est possible, la méthodologie cherche à décider si l'entité qui correspond au type de la question répond bien à cette question grâce à une représentation logique et une justification logique basées sur la connaissance du monde apportée par *WordNet*, ainsi qu'une résolution de coréférence au niveau du paragraphe considérée comme rare dans cette fenêtre. Cette partie

logique de la méthodologie est peu détaillée et peu convaincante, peut-être à cause de la nature commerciale du système, dont nombre de spécificités et de fonctionnements restent confidentiels.

1.3 Conclusion

Suite à un examen de différentes méthodologies appartenant aux disciplines les plus exigeantes vis-à-vis de la précision d'une information désirée, nous pouvons tirer certaines conclusions. Tout d'abord, dès lors qu'il s'agit de prendre connaissance du contenu d'un texte, l'analyse linguistique semble inévitable, même si certaines approches n'y font qu'un appel très marginal. Les systèmes les plus récents s'essaient d'ailleurs à des analyses linguistiques de plus haut niveau, y intégrant la syntaxe et surtout la sémantique. Ensuite, le principe appliqué pour détecter une information donnée est systématiquement de donner à cette information le plus grand nombre de présentations différentes et de comparer ces présentations avec le contenu des textes. L'extraction d'information constitue pour cela des listes de patrons ou de tableaux, tandis que la discipline de question-réponse y préfère l'expansion de requête. Dans les deux cas, des lexiques ou bases de connaissances sont fréquemment exploités.

Par ailleurs, cet examen nous a permis d'identifier les besoins que le domaine peut avoir d'une structure sémantique informationnelle constituée à partir d'une base textuelle. En fonction des approches étudiées, il s'agit d'effectuer une identification lexicale, morpho-lexicale, syntaxique voire sémantique de l'information. Il s'agit également de tenir compte d'une éventuelle structure textuelle (textes structurés ou semi-structurés) et de pouvoir en rendre compte.

Dès lors, l'analyse que nous allons faire du texte devra prendre en compte ces attentes, tout en se montrant capable de fournir une base à un enrichissement considérable de l'information présente. Par ailleurs, la structure devra se révéler accessible pour son interrogation. Le chapitre prochain va décrire les outils d'analyse qui permettront d'identifier l'information présente dans la base documentaire avant d'effectuer son enrichissement.

Chapitre 2

Les outils d'analyse textuelle

2.1 Introduction

L'approche de la problématique de la gestion de l'information que nous avons choisie, contrairement aux méthodes habituelles des domaines de l'extraction d'information ou de question-réponse, est centrée sur l'analyse des textes plutôt que sur l'étude des données contenues dans la requête, dont le contexte est limité. Le volume des documents à étudier est de ce fait bien plus important que les énoncés des requêtes ou les quelques fragments de textes proposés comme réponse à la requête.

Par ailleurs, nous avons fait le choix de rejeter les méthodes statistiques au profit d'une approche linguistique du texte. Les informations contenues dans la base documentaire doivent donc être identifiées au cours d'une analyse linguistique. Ces données sont de trois types : morpho-lexical (les mots), syntaxique (les relations syntaxiques entre les mots) et lexico-sémantique (le sens des mots).

La démarche de la structuration de l'information est composée de deux phases étroitement interconnectées. Il s'agit d'abord de l'identification de l'information contenue dans les textes, qui consiste principalement en une série d'analyses linguistiques. Vient ensuite la phase d'enrichissement durant laquelle les résultats de l'analyse sont utilisés pour sélectionner l'information lexicale destinée à enrichir la structure informationnelle. La figure 2.1 page suivante illustre l'architecture du système de structuration de l'information d'une base textuelle. On peut y voir combien les deux phases du processus (*identification* de l'information par l'analyse et *enrichissement* des données identifiées) sont imbriquées l'une dans l'autre.

Ce chapitre est consacré aux outils d'analyse permettant de traiter l'information de la base documentaire. Pour identifier l'information contenue

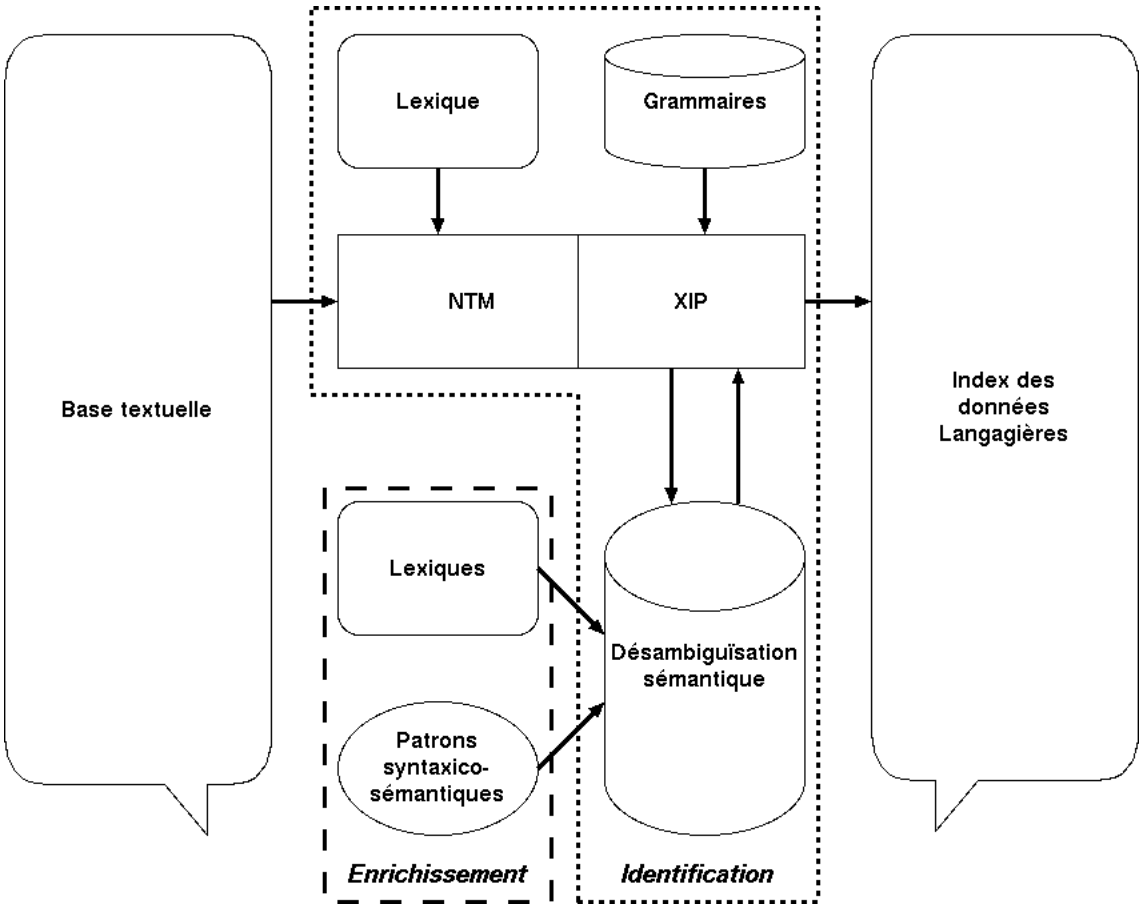

FIG. 2.1 – Schéma de l'architecture du système de structuration des documents.

dans les textes, nous avons besoin d'**analyseurs robustes** capables d'identifier les mots (segmentation et analyse morphologique), les relations (analyse syntaxique) et le sens des mots (désambiguïsation sémantique). Comme la désambiguïsation sémantique est appelée à déterminer les enrichissements de la structure informationnelle, nous étudions plus en profondeur les différentes méthodes existantes avant de décrire la méthodologie qui nous semble la plus adaptée à notre démarche.

2.2 La chaîne d'analyse morpho-syntaxique : *NTM* et *XIP*

Lorsque l'on s'attache à l'étude de la structure sémantique d'un texte, plusieurs possibilités transparaissent pour gérer l'information. Dans le même temps, certains impératifs émergent, qui concernent l'utilisation d'une ressource lexicale pour identifier les unités lexico-sémantiques présentes et un système d'indexation pour localiser ces unités. Les choix reposent essentiellement sur l'utilisation que l'on fait de l'information de ce dictionnaire et sur la méthode d'analyse de l'articulation des concepts les uns par rapport aux autres.

Pour agencer les entités les unes vis-à-vis des autres et faire évoluer leur sens en fonction des rapports qu'elles entretiennent entre elles, nous avons choisi d'effectuer une analyse morpho-syntaxique du texte. Cette analyse permet en effet de distinguer de nombreuses relations qui unissent les unités qui composent le texte. La chaîne d'analyse passe par une identification des lexèmes lors de la phase de segmentation du texte (*tokenization*) et une analyse morphologique de chacun des mots – ces tâches sont effectuées par un module appelé *NTM* (*Normalizer – Tokenizer – Morphological analyzer*) ; ensuite par un outil d'analyse syntaxique robuste (*XIP* : *Xerox Incremental Parser*) dont les grammaires permettent d'effectuer une désambiguïsation morpho-syntaxique des propositions de l'analyseur morphologique ainsi que de construire des groupes syntaxiques et de définir les relations syntaxiques entre les têtes de ces groupes. Ces opérations sont effectuées par l'entremise de règles spécifiques constituées en grammaires, qui s'appliquent incrémentalement en s'appuyant sur des traits morpho-syntaxiques et sur la disposition contextuelle des unités qui constituent les énoncés. Ce moteur d'analyse est dit robuste, ce qui implique qu'un succès n'est pas nécessaire dans l'analyse de tous les constituants de l'énoncé pour que l'analyseur fournisse un résultat.

Ces deux outils sont développés à XRCE (Xerox Research Centre Europe) Grenoble (XRCE) et fonctionnent actuellement en parfaite symbiose puisqu'elles ont été intégrées au sein d'une seule application qui permet de recevoir un texte (en français dans notre cas, mais l'anglais est également couvert) brut ou formaté et d'en retourner une analyse morphologique et syntaxique. Étant donné que *XIP* est un outil robuste, il présente une analyse dans tous les cas, qui sera partielle lorsqu'il n'a pu parvenir à un résultat. Nous décrivons ici ces deux étapes d'analyse et les outils qui les accomplissent.

2.2.1 Normalisation et analyse morphologique

La première étape de cette analyse morpho-syntaxique consiste à segmenter le texte présenté en lexèmes et à en donner les caractéristiques morphologiques possibles. Ces deux tâches sont effectuées simultanément par *NTM*, un module conçu par Salah Aït-Mokhtar [Aït-Mokhtar, 1998, Trouilleux, 2001] qui repose sur la technologie des **automates à états finis**. Deux **transducteurs** ou plus contiennent les données utiles à l'analyse que l'on veut réaliser et fonctionnent en parallèle pour produire le résultat. Ce fonctionnement simultané de deux automates permet d'effectuer dans un même temps la segmentation et l'analyse morphologique.

Un de ces transducteurs est le normaliseur (*normalizer*), dont le rôle est de transmettre une chaîne de caractères normalisés aux autres transducteurs. La normalisation des caractères se fait selon des règles propres à la langue traitée et aux exigences de l'utilisateur. Pour le français, elles portent essentiellement sur la casse et sur l'accentuation, ainsi que sur le formatage typographique du texte présenté en entrée. Dans notre cas, la base textuelle est au format XML et de nombreux caractères sont présents dans les fichiers sous la forme d'entités. Par exemple, le caractère *é* est transcrit *é* dans le texte, mais il ne peut être reconnu directement dans le lexique. Le rôle du normaliseur est de transmettre cette entité sous la forme d'un caractère standard aux autres transducteurs.

Ce second[1] transducteur contient les données lexicales permettant de reconnaître chaque lexème et d'en donner toutes les informations d'ordre morphologique correspondantes, ainsi que toute autre information qui y aurait été ajoutée. Il reçoit caractère par caractère une chaîne de caractères normalisés à laquelle il tente de faire correspondre les données dont il dispose pour obtenir les analyses possibles. Cependant, comme *NTM* obéit au principe de la chaîne la plus longue (*longest match*), les analyses obtenues sont susceptibles d'être surclassées ensuite par une correspondance d'une chaîne plus longue. En effet, c'est toujours l'analyse conforme à la chaîne la plus longue qui prévaudra, tandis que les précédentes propositions, plus courtes, ne seront pas maintenues. Cette stratégie, qui permet d'identifier les unités lexicale composées de lexèmes plus courts, se révèle très efficace dans la pratique [Aït-Mokhtar, 1998].

Il faut aussi noter l'importance des séparateurs de mots, dont certains – tels que l'espace, les ponctuations, le parenthésage ou les guillemets – sont définis par défaut, mais qui restent paramétrables par l'utilisateur. Ils revêtent une grande importance lors de la mise en application de l'analyse.

[1] Il n'est le second que dans notre description. En effet, tous les transducteurs dont *NTM* tire parti effectuent simultanément leur tâche.

En effet, le fonctionnement standard [2] de *NTM* prévoit qu'un découpage et une analyse ne peuvent être validés que si un séparateur de mots limite la chaîne proposée et reconnue. De la sorte, une suite de caractères reconnus mais qui n'est pas limitée par un séparateur n'aboutira pas à une analyse valide.

```
[\#\#]> echo "Pour son \&eacute;l\&egrave;ve, il
        commence..." | ntm
```

Analyses morphologiques produites par *NTM* :

```
Pour                    pour        +Masc+InvPL+Noun
Pour                    pour        +Prep+PREP

son                     son         +PP3S+InvGen+SG+Poss+Det
son                     son         +Masc+SG+Noun+NOUN_SG

&eacute;l&egrave;ve     élever      +se+contreSN+avoir+SubjP+SG+P1+Verb
&eacute;l&egrave;ve     élever      +se+contreSN+avoir+SubjP+SG+P3+Verb
&eacute;l&egrave;ve     élever      +se+contreSN+avoir+IndP+SG+P1+Verb
&eacute;l&egrave;ve     élever      +se+contreSN+avoir+IndP+SG+P3+Verb
&eacute;l&egrave;ve     élever      +se+contreSN+avoir+Imp+SG+P2+Verb
&eacute;l&egrave;ve     élève       +deSN+InvGen+SG+Noun

,                       ,           +CM

il                      il          +Nom+Masc+SG+P3+PC

commence                commencer   +avoir+parSN+Imp+SG+P2+Verb
commence                commencer   +avoir+parSN+IndP+SG+P1+Verb
commence                commencer   +avoir+parSN+IndP+SG+P3+Verb
commence                commencer   +avoir+parSN+SubjP+SG+P1+Verb
commence                commencer   +avoir+parSN+SubjP+SG+P3+Verb

...                     ...         +PUNCT
```

FIG. 2.2 – Propositions d'analyse morphologique d'un énoncé par *NTM*.

L'exemple que nous présentons dans la figure 2.2 illustre les différentes informations que l'analyseur morphologique fournit en sortie. Chaque ligne correspond à une hypothèse d'analyse pour une unité lexicale présentée, et un interligne distingue les hypothèses d'analyse pour chaque unité. Chaque ligne comprend trois champs : le premier contient l'unité lexicale telle qu'elle se présente dans le texte, le deuxième présente la lemmatisation de cette unité

[2] Il est possible de paramétrer *NTM* pour forcer la reconnaissance de préfixes. Dans ce cas, les préfixes pourront être analysés sans pour cela être délimités par un séparateur.

correspondant à l'analyse morphologique affichée sous forme d'étiquettes (séparées par `+`) dans le troisième champ, accompagnée de toute autre information dont nous parlions plus haut.

Au travers de l'exemple, on peut voir l'action du système de normalisation des caractères : par exemple, la majuscule de début de phrase n'est pas conservée lors de l'analyse de *Pour* et sa lemmatisation en `pour`, tandis que les entités XML `&eacute;` et `&egrave;` sont résolues respectivement par `é` et `è` pour permettre une analyse ambiguë entre les lemmes *élever* et *élève*.

Cet exemple nous permet aussi d'illustrer le type d'information que fournit l'analyseur morphologique dans le troisième champ. La dernière étiquette à droite indique la catégorie grammaticale à laquelle appartient l'unité lexicale courante (`Adj` pour adjectif, `Adv` pour adverbe, `Det` pour déterminant, `Noun` pour nom, `PC` pour pronom clitique, `Prep` pour préposition, `Verb` pour verbe, etc.).

Ainsi, la forme normalisée *élève*, par exemple, peut s'analyser comme un verbe à travers la forme lemmatisée *élever* ou comme un nom, *élève*. Les étiquettes `P1`, `P2` et `P3` contiennent l'information relative à la personne, les étiquettes `SG` et `PL` celles relatives au nombre et les étiquettes `Masc`, `Fem` et `InvGen` (pour invariable en genre) celles relatives au genre.

L'information temporelle et modale des verbes est spécifiée par les étiquettes `SubjP` (subjonctif présent), `IndP` (indicatif présent) et `Imp` (impératif), complétée par la détermination de l'auxiliaire avec lequel le verbe construit ses formes composées actives. Les pronoms reçoivent une information casuelle : au `Nom` de nominatif correspond un `Acc` pour accusatif. Cette information casuelle, qui n'est pas toujours pertinente en français, pourra rester inexploitée.

Les déterminants sont susceptibles de recevoir aussi une information relative à leur caractère défini (`Def` pour défini et `Indef` pour indéfini). Le cas de *des*, préposition contractée avec l'article *le*, est particulier car il s'agit de fournir l'analyse de deux lexèmes condensés en une seule unité lexicale. Aussi l'étiquette `Prep=le` indique-t-elle que la préposition lemmatisée en *de* est contractée avec le lexème *le*, lequel est un article défini (`Def`). On voit dans l'exemple que le cas de *des* article indéfini et celui de *des* préposition contractée sont étudiés en parallèle, aucun des deux n'excluant l'autre.

Dans son état original, le dictionnaire du français encode une seule information sortant du cadre strict de l'analyse morphologique. Il s'agit de données de **sous-catégorisation**. Dans notre exemple, on ne retrouve que dans le cas du verbe et du nom cette information relative aux compléments. Le verbe *élever* comporte les étiquettes `se`, qui autorise sa pronominalisation, `SN`, qui marque la transitivité (sous forme d'un syntagme nominal), et `contreSN`, qui signale la présence d'un complément prépositionnel optionnel composé de la préposition *contre* et d'un syntagme nominal. De même, le nom *élève* comporte une étiquette proposant un complément prépositionnel composé de la préposition *de* et d'un syntagme nominal.

Cet outil présente deux principaux atouts pour la tâche que nous avons à effectuer. D'une part, il est relativement aisé d'insérer dans le transducteur contenant le lexique une information supplémentaire liée au lexique qui pourra être exploitée avec profit par les grammaires de l'analyseur syntaxique. De l'autre, le caractère paramétrable de la normalisation permet de

l'adapter aux besoins liés au format XML des textes de la base documentaire que nous avons à notre disposition.

2.2.2 Analyse syntaxique

L'analyse linguistique d'un texte visant à sa compréhension au moins partielle requiert un examen élaboré de sa structure syntaxique, car celle-ci permet de mettre en relation les éléments constitutifs du sens de ce texte. Les outils syntaxiques qui fonctionnent efficacement sur de grandes bases de données textuelles francophones ne sont cependant pas très nombreux. Nous disposons toutefois de deux d'entre eux à XRCE (Xerox Research Centre Europe), ***IFSP*** et *XIP*, décrits dans les sections suivantes.

IFSP

Le premier analyseur s'appelle *IFSP* pour *Incremental Finite-State Parser* [Aït-Mokhtar et Chanod, 1997a, Aït-Mokhtar et Chanod, 1997b]. Comme l'indique son nom, il exploite, à l'instar de *NTM*, la technologie des transducteurs à états finis. Ces transducteurs encodent une information contextuelle constituant les règles permettant d'obtenir les groupes et les relations syntaxiques. Ces règles s'appliquent incrémentalement, c'est-à-dire qu'elles sont successivement mises en œuvre pour obtenir un résultat, les résultats de l'une servant de point de départ à la suivante. De la sorte, toute analyse atteint les possibilités maximales de l'analyseur sans qu'un échec puisse contrarier son fonctionnement. *IFSP* possède deux niveaux de fonctionnement. Dans un premier temps, à partir du résultat d'une analyse morphologique préalablement désambiguïsée fournie en entrée, l'analyseur syntaxique construit les **groupes syntaxiques minimaux** (*chunks*). Vient ensuite l'extraction des dépendances internes et externes à ces groupes syntaxiques.

Les groupes syntaxiques minimaux correspondent aux groupes de la grammaire traditionnelle (groupes nominaux, verbaux et prépositionnels), à cette différence près qu'ils sont tronqués de leur partie droite au-delà de la tête du groupe. Ces groupes minimaux sont nominaux (`NP`), verbaux (`v`), adjectivaux (`AP`) ou prépositionnels (`PP`). Tous peuvent être incorporés à l'intérieur d'une phrase minimale (`SC`) qui constitue l'unité de référence phrastique de l'analyseur. La fenêtre de travail quant à elle est la phrase, comprise entre deux ponctuations fortes.

L'exemple 2.3 page ci-contre illustre le découpage syntaxique de la phrase en **syntagmes minimaux**. En particulier, sous sa forme de syntagme minimal, le groupe nominal traditionnel *son homologue britannique* est limité à droite par sa tête *homologue*, tandis que l'épithète *britannique* constitue un syntagme adjectival, au même titre que l'attribut *content*. Il est cependant

possible de reconstituer le groupe nominal traditionnel grâce à la dépendance construite entre *homologue* et *britannique* (`PADJ(homologue,britannique)`). On notera la différence de traitement avec *Le nouveau ministre*, où l'épithète fait partie du syntagme minimal car elle est placée devant la tête du syntagme.

Les dépendances syntaxiques établissent des liens entre deux ou trois unités lexicales selon le type de dépendance, qui constituent dès lors les arguments de la dépendance. Lorsque ces dépendances concernent non pas des lexèmes, mais des groupes syntaxiques minimaux, c'est la tête du groupe syntaxique qui fait office d'argument, de telle sorte que toutes les dépendances extraites par *IFSP* associent formellement des lexèmes entre eux.

```
[##]> echo "Le nouveau ministre du budget est
content de recevoir son homologue britannique." |
ifsp -l french

[SC [NP Le nouveau ministre NP]/SUBJ [PP du budget
PP] :v est SC] [AP content AP] [v de recevoir v]
[NP son homologue NP]/OBJ [AP britannique AP]

SUBJ(ministre,recevoir)
SUBJ(ministre,être)
DOBJ(recevoir,homologue)
ADJ(nouveau,ministre)
PADJ(homologue,britannique)
ATTR(ministre,content)
NNPREP(ministre,de=le,budget)
```

FIG. 2.3 – Exemple d'analyse d'un énoncé par *IFSP*.

L'exemple 2.3 reproduit le résultat d'une analyse effectuée par *IFSP*. Cette analyse comporte deux parties : le découpage de l'énoncé en syntagmes minimaux (`SC`, `NP`, `PP`, `v`, `AP`) et la construction des dépendances syntaxique. Les dépendances ainsi construites indiquent que *ministre* est le sujet (`SUBJ`) de *recevoir* et de *être*, que *homologue* est l'objet direct (`DOBJ`) de *recevoir*, que *britannique* est l'épithète du nom précédent (`PADJ`) *homologue* et *nouveau* l'épithète du nom suivant (`ADJ`) *ministre*, que *content* est l'attribut (`ATTR`) de *ministre* et que *budget* est rattaché au nom *ministre* à travers la préposition *du* (`NNPREP`).

Malgré les grandes qualités de cet outil, tant par la robustesse que par la rapidité ou le niveau des résultats obtenus, il présente le défaut majeur de la rigidité. En effet, une grammaire sous forme de transducteurs est complexe à modifier, plus encore à enrichir de nouvelles dépendances. Un autre défaut

essentiel de cet analyseur pour l'application que nous voulons en faire est qu'il n'est pas capable de gérer les traits que nous désirons associer dans les dépendances aux lexèmes qui en sont les arguments. De fait, ces traits sont un élément prépondérant de la sémantique lexico-syntaxique que nous mettons en œuvre pour aboutir à une compréhension des documents.

XIP

Au-delà des possibilités limitées de *IFSP*, *XIP* (*Xerox Incremental Parser*) reprend à son compte le principe de l'incrémentalité qui conférait efficacité et robustesse à *IFSP*. *XIP* y ajoute cependant une souplesse d'utilisation et des fonctionnalités qui lui permettent d'envisager l'inclusion d'éléments de sémantique à son analyse [Roux, 1999, Aït-Mokhtar et al., 2002, Hagège et Roux, 2002]. Nous allons voir que ces possibilités permettent d'envisager une exploitation de *XIP* dans une perspective syntaxico-sémantique intéressante.

```
[##]> echo "Le nouveau ministre du budget est
content de recevoir son homologue britannique." |
xip

0>GROUPE{SC{NP{Le AP{nouveau} ministre} PP{du
NP{budget}} FV{est}} AP{content} IV{de recevoir}
NP{son homologue} AP{britannique} .}

SUBJ[NOUN](est,ministre)
VARG[DIR](recevoir,homologue)
VARG[ADJ,SPRED](est,content)
NMOD[RIGHT,ADJ](homologue,britannique)
NMOD[LEFT,ADJ](ministre,nouveau)
NMOD[NOUN,INDIR](ministre,du,budget)
NMOD[ADJ,SPRED](ministre,content)
ADJARG[INF,INDIR](content,recevoir)
```

FIG. 2.4 – Analyse par *XIP* correspondant à celle par *IFSP* (voir figure 2.3).

Dans l'exemple 2.4, on peut voir le résultat d'une analyse syntaxique tel que *XIP* la restitue. Comme pour *IFSP*, l'affichage du résultat comporte deux parties distinctes. La première partie montre le découpage de l'énoncé en syntagmes minimaux semblables à ceux que réalise *IFSP* (`SC`, `NP`, `AP`, `PP`, mais `FV` pour le groupe verbal et `IV` pour la base d'une infinitive). La seconde partie exprime les dépendances syntaxiques. Elles présentent la particularité, par rapport à celles qu'extrait *IFSP*, d'être peu nombreuses mais

de comporter des traits qui permettent de les distinguer entre elles. Le premier mot de la dépendance en indique la nature (`SUBJ` pour sujet, `VARG` pour argument d'un verbe, `NMOD` pour modifieur de nom et `ADJARG` pour argument d'un adjectif). Les mots entre crochets sont des traits qui portent sur la nature de la dépendance [3]. Les arguments des dépendances sont placés entre parenthèses. La principale différence directement détectable avec *IFSP* repose dans l'utilisation de traits pour distinguer les différentes dépendances de nature pourtant semblable. Ainsi, une même dépendance `NMOD` décrit la dépendance qui unit un adjectif et un nom, que cet adjectif en soit attribut (trait `SPRED`) ou épithète (trait `LEFT` si l'adjectif précède le nom et `RIGHT` si l'adjectif le suit). L'étude de cet analyseur permet de détecter d'autres fonctionnalités de *XIP*.

XIP est d'abord composé d'un moteur d'analyse développé par Claude Roux et ensuite d'une grammaire propre à la langue du texte analysé. Le moteur exploite la grammaire constituée d'un ensemble de règles explicites. La grammaire du français disponible dans le formalisme de *XIP* a été écrite par Jean-Pierre Chanod et Salah Aït-Mokhtar. Elle dérive donc naturellement de celle de *IFSP*. Cependant, la distinction entre moteur et grammaire permet d'enrichir ou de modifier à volonté cette dernière dans la limite du formalisme proposé par le moteur.

Au travers de sa grammaire décrivant la structure syntaxique des phrases en français, *XIP* vise comme *IFSP* à fournir une analyse syntaxique finale en termes de dépendances. Ce résultat ne peut être atteint que par l'intermédiaire d'une phase de constitution des groupes syntaxiques minimaux. Toutefois, contrairement à *IFSP*, l'étape d'analyse s'effectue désormais dans le cadre de la construction d'un arbre d'analyse partiel constitué de nœuds. Par ailleurs, en ce qui concerne l'étiquetage morphologique et donc le traitement de l'ambiguïté catégorielle, la grammaire française actuellement assignée à *XIP* préfère l'usage de règles contextuelles de désambiguïsation catégorielle au modèle statistique, dit « **modèle de Markov caché** » (*Hidden Markov model*, HMM). *IFSP* exploitait naguère les choix de ce modèle statistique pour construire ses syntagmes minimaux et établir ses relations syntaxiques.

L'étiquetage morphologique repose donc sur un ensemble de règles de désambiguïsation catégorielle qui constituent la première étape de l'analyse de *XIP* dès lors que la normalisation, le découpage en mots et l'analyse morphologique ont eu lieu au travers de *NTM* (ou de tout autre outil d'analyse morphologique qu'on lui préférerait). Lors de l'analyse morphologique,

[3] `NOUN` indique que l'argument de droite est un nom, `ADJ` que c'est un adjectif et `INF` que c'est un verbe à l'infinitif, `DIR` indique que la dépendance est directe et `INDIR` qu'elle est indirecte, `RIGHT` indique que le dernier argument de la dépendance se situe dans le contexte droit du premier et `LEFT` qu'il se situe dans son contexte gauche, `SPRED` signifie que l'argument adjectival est attribut du nom dans le cas d'un `NMOD`, ou du sujet du verbe dans le cas d'un `VARG`.

chacune des unités lexicales reçoit autant d'étiquettes morphologiques que son actualisation dans le texte permet d'en trouver. Ces différentes analyses permettent de constituer une classe d'ambiguïté pour chaque unité lexicale. Les règles de désambiguïsation catégorielle prennent en compte les classes d'ambiguïté de chacune des unités lexicales et tentent de réduire ces classes en confrontant chaque proposition aux classes d'ambiguïté du contexte de l'unité analysée.

Nous avons préféré la stratégie à base de règles pour effectuer la désambiguïsation catégorielle à la méthode statistique basée sur un **HMM**. Ce choix a été motivé par la possibilité, dans les règles, d'exploiter certains traits morphologiques en plus des catégories que les méthodes statistiques utilisent exclusivement. De la sorte, tous les éléments morpho-linguistiques de l'unité lexicale en cours d'analyse ainsi que ceux de son contexte peuvent être exploités pour parvenir à une meilleure désambiguïsation catégorielle. La méthode statistique fondée sur le HMM ne propose qu'une fenêtre contextuelle très limitée de deux mots pour effectuer un choix, et elle ne fonde ce choix que sur l'information catégorielle de sa fenêtre contextuelle. Elle affiche donc des possibilités restreintes par rapport à une méthode à base de règles.

Dans l'exemple 2.5 page suivante, nous indiquons les propositions de l'analyseur morphologique et la traduction sous forme de traits qu'en fait *XIP* afin de manipuler tous les éléments informatifs dont il a besoin pour effectuer son analyse. La règle de désambiguïsation catégorielle permet de restreindre partiellement l'ensemble des propositions d'analyse. En effet, à l'intérieur d'une classe d'ambiguïté dans laquelle sont présents un verbe (`verb`) et un nom (`noun`), elle prescrit de ne conserver que les propositions (quelles qu'elles soient, comme le permet le `?`) ne dénotant pas un trait verbal (`verb :~`) lorsque le contexte gauche contient un déterminant (`det`) non ambigu, c'est-à-dire sans trait dénotant un pronom (`pron :~`), une préposition (`prep :~`) ou un adjectif (`adj :~`). Ce déterminant peut éventuellement [4] être suivi d'un ou plusieurs adjectifs (`adj`) dont un trait (`adj2`) indique que cet adjectif peut précéder le lexème dont il est épithète.

L'énoncé présenté permet l'application de la règle de désambiguïsation, car le contexte correspond à sa condition : un déterminant non ambigu (*sa*) puis des adjectifs (*première* et *grande*) sans limitation de nombre (`*`) précèdent le lexème *mesure* dont la classe d'ambiguïté contient les catégories *nom* et *verbe*. La règle sélectionne dès lors les propositions qui ne dénotent pas le trait `verb` (`verb :~`). Une seule proposition d'analyse est donc maintenue [5].

[4] L'astérisque indique, comme dans les expressions régulières, que l'expression qui la précède n'est pas obligatoire et qu'elle peut être présente un nombre illimité de fois.

[5] Dans le cas où la désambiguïsation catégorielle n'a pu aboutir au maintien d'une seule proposition, l'ensemble des propositions restantes est conservé et chacune de ces propositions est utilisée pour les étapes ultérieures. Celle qui permet la meilleure analyse

Résultat de l'analyse morphologique par *NTM* :

```
[##]> echo "mesure" | ntm

mesure  mesurer +SubjP+SG+P1+Verb
mesure  mesurer +SubjP+SG+P3+Verb
mesure  mesurer +Imp+SG+P2+Verb
mesure  mesurer +IndP+SG+P1+Verb
mesure  mesurer +IndP+SG+P3+Verb
mesure  mesure  +Fem+SG+Noun
```

Traduction de l'analyse morphologique dans le formalisme de traits *XIP* :

```
verb[subj:+,pre:+,sg:+,p1:+,verb:+]
verb[subj:+,pre:+,sg:+,p3:+,verb:+]
verb[imp:+,sg:+,p2:+,verb:+]
verb[ind:+,pre:+,sg:+,p1:+,verb:+]
verb[ind:+,pre:+,sg:+,p3:+,verb:+]
noun[fem:+,sg:+,noun:+]
```

Règle de désambiguisation catégorielle :

```
noun,verb = |det[pron:~,prep:~,adj:~],adj*[adj2]| ?[verb:~].
```

Énoncé sur lequel la règle s'applique :

Sa première grande **mesure** contre les disciples de Jésus concerne l'enseignement.

FIG. 2.5 – Construction et fonctionnement d'une règle de désambiguïsation catégorielle.

L'utilisation de telles règles de désambiguïsation catégorielles permet de dégager certaines différences par rapport à la méthode statistique HMM. Nous avons déjà mentionné une exploitation du contexte plus adaptée au traitement linguistique qu'une simple fenêtre, car la règle de désambiguïsation permet à la fois de sélectionner dans le contexte les éléments importants pour effectuer un choix catégoriel et de faire varier l'étendue du contexte pris en considération en fonction des besoins (notamment grâce aux expressions de généralisation `?`, `*` et `+` issues des quantificateurs des expressions régulières). Un autre avantage des règles de désambiguïsation provient de l'exploitation de différents traits linguistiques là où seules les catégorie grammaticale du contexte étaient utilisées par le HMM. Enfin, la méthode HMM utilisait un

est finalement conservée.

corpus d'apprentissage pour établir une matrice qui permettait d'effectuer la désambiguïsation catégorielle. En cas de mauvais fonctionnement, il fallait constituer un nouveau corpus et reconstruire une matrice. L'utilisation de règles permet l'adjonction ou la correction aisée d'une ou plusieurs règles lorsque des erreurs sont constatées.

Une des particularités de *XIP*, dans le contexte de la grammaire utilisée, est de construire un arbre syntaxique partiel pour aboutir aux groupes syntaxiques minimaux nécessaires pour l'extraction des dépendances. Cet arbre syntaxique est composé de nœuds qui peuvent être de deux types : lexicaux ou non lexicaux. Ces nœuds sont porteurs de toute l'information linguistique liée à l'analyse effectuée par *XIP* excepté les dépendances. En effet, il n'entre pas dans les propriétés des nœuds de décrire les relations qu'ils entretiennent les uns avec les autres.

Les nœuds lexicaux sont ceux qui dominent immédiatement les feuilles de l'arbre syntaxique, c'est-à-dire les unités lexicales. L'information dont ils sont porteurs correspond à la catégorie grammaticale issue de l'analyse morphologique puis de la désambiguïsation catégorielle de l'unité lexicale correspondante. Il s'agit donc d'un simple étiquetage morpho-syntaxique des feuilles intégré à l'intérieur de la structure arborescente.

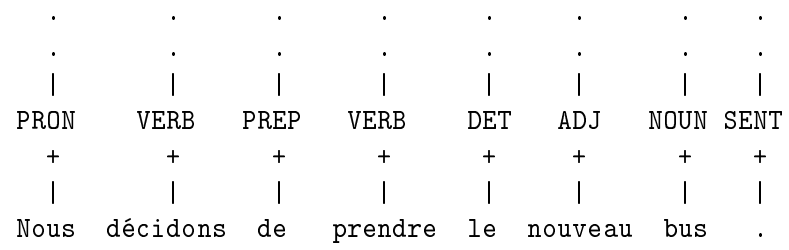

FIG. 2.6 – Étiquetage des feuilles par les nœuds lexicaux dans l'arbre syntaxique partiel.

Dans le cadre de la grammaire française, les groupes syntaxiques ou syntagmes construits sont dits minimaux. Il s'agit en effet de construire un arbre syntaxique qualifié de partiel. Cela implique qu'il ne vise pas la description syntaxique de l'ensemble de la structure phrastique, mais uniquement son découpage en unités essentielles, qui sont les syntagmes minimaux. Comme dans *IFSP*, un groupe syntaxique minimal correspond à un groupe syntaxique traditionnel dont on aurait retranché la partie située à droite de la tête [6]. Le choix de construire des syntagmes partiels comme c'est le cas pour les groupes syntaxiques minimaux permet, lors de la construction de l'arbre syntaxique partiel, de ne pas préjuger des relations syntaxiques entre la tête d'un de ces groupes et une unité lexicale qui la suit dans la phrase. L'établis-

[6] [Grevisse et Goosse, 1991] [§270] appelle la tête « noyau » du groupe syntaxique.

sement d'éventuelles relations sera effectué, le cas échéant, lors de la phase d'extraction des dépendances. Un noyau non lexical correspond à l'étiquette syntaxique du groupe syntaxique minimal qu'il domine immédiatement.

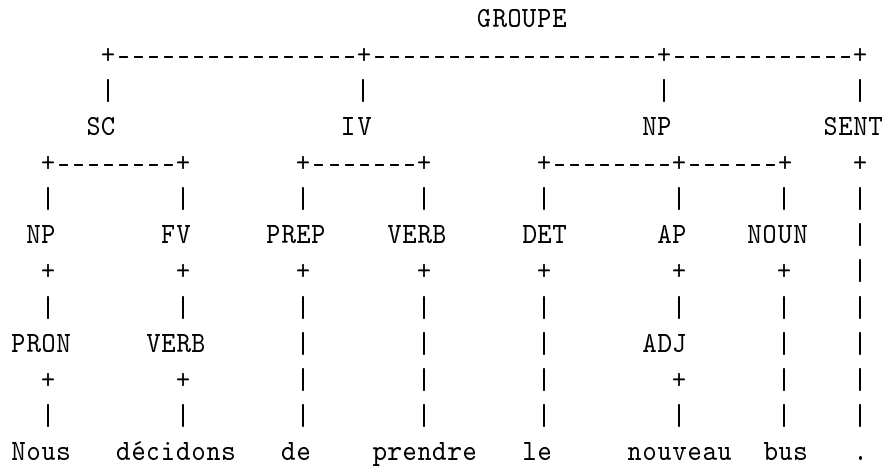

FIG. 2.7 – Exemple d'un arbre syntaxique partiel.

Une fois construit l'arbre syntaxique partiel, la tâche de *XIP* consiste à décrire la structure syntaxique de la phrase sous forme de dépendances. Il s'agit là d'un aspect particulièrement important de l'analyse syntaxique, les relations entre syntagmes et entre lexèmes constituant la base des liens sémantiques que nous cherchons à établir entre les concepts sémantiques dont ils sont porteurs. Les dépendances syntaxiques générées par *XIP* sont les étiquettes des relations établissant un lien entre différents nœuds de l'arbre syntaxique partiel. Selon qu'il s'agit d'un nœud lexical ou non lexical, la dépendance affichera donc un lexème pour lui-même[7] ou bien en tant que tête du syntagme minimal[8] représenté par le nœud.

Dans le formalisme de *XIP*, les dépendances syntaxiques sont dès lors exprimées selon le format :

```
PREDICAT(Arg1,Arg2,...,ArgN)
```

où `Arg1,Arg2,...,ArgN` correspondent à des nœuds de l'arbre syntaxique partiel entre lesquels est établie la dépendance dont le nom est `PREDICAT`. Étant donné que les arguments de la relation décrite sont des nœuds qui peuvent être lexicaux ou non lexicaux, le formalisme de *XIP* les représente, pour des raisons de lisibilité, par le nœud terminal le plus à droite (la tête) dans ceux qui sont dominés par chaque nœud constituant un argument. Dans notre exemple (*cf.* figures 2.7 et 2.8 page suivante, le groupe *le nouveau bus* constitue un syntagme minimal `NP`, objet direct (`VARG[DIR]`) de *prendre*,

[7] L'affichage du lexème représente uniquement ce lexème.
[8] L'affichage du lexème représente le syntagme minimal dont ce lexème est la tête.

ce que *XIP* exprime par une dépendance `VARG[DIR](prendre,bus)`. C'est donc une forme lexicale de surface qui fait office d'argument pour la relation, ce qui permet de grouper dans une même expression deux unités lexicales syntaxiquement liées dans l'énoncé. Les résultats qui apparaissent dans ces exemples ne sont toutefois qu'un affichage. Lorsque *XIP* manipule une unité lexicale ou une dépendance, l'ensemble des traits qui y sont associés sont présents et peuvent être utilisés et exploités.

Notons encore que le nombre d'arguments d'une dépendance n'est pas limité par le formalisme de *XIP*. Les dépendances les plus fréquentes possèdent deux ou trois arguments, indiquant ainsi une relation entre ces arguments. La formulation d'une dépendance unaire est toutefois possible, et même fréquente dans la grammaire du français. Cette particularité permet de définir des dépendances informatives liées à une seule unité lexicale, comme on le verra notamment dans la description de l'interrogation de la structure sémantique informationnelle, avec la constitution d'une dépendance `FOCUS` (*cf.* section 6.2.2 page 187).

```
 27> echo "Nous décidons de prendre le nouveau bus." | xip
SUBJ(décidons,Nous)
SUBJ(prendre,Nous)
VARG[DIR](prendre,bus)
VARG[INF,INDIR](décidons,prendre)
NMOD[LEFT,ADJ](bus,nouveau)
PREPOBJ[INF](de,prendre)
0>GROUPE{SC{NP{Nous} FV{décidons}} IV{de prendre} NP{le
AP{nouveau} bus} .}
```

FIG. 2.8 – Dépendances extraites par *XIP* : les feuilles remplacent les nœuds.

Outre les réelles qualités de *XIP* dont nous avons parlé précédemment, à savoir la possibilité de modifier et d'enrichir la grammaire présente – ou de la remplacer – selon les besoins ressentis par l'utilisateur, certaines autres particularités propres à cet outil sont décisives face aux arguments qui doivent mener au choix d'un analyseur syntaxique. Notamment, la possibilité de manipuler des traits au cours de l'analyse est un avantage de premier plan dans la perspective de l'utilisation sémantique que nous désirons en faire. En effet, non seulement le système de gestion des traits permet de dépasser la syntaxe pour tendre vers la sémantique, mais de plus il est loisible d'étendre indéfiniment ou presque la variété de ces traits.

Les traits de *XIP* ont la forme de couples `attribut :valeur`. Ils peuvent être attachés soit à une forme lexicale (une feuille de l'arbre syntaxique partiel), soit à un nœud de l'arbre syntaxique, soit encore à une dépendance

portant sur un ou des nœuds de cet arbre [9]. De la sorte, un trait déterminant un mot, un nœud ou une dépendance peut recevoir des modifications en cours d'analyse en fonction de son environnement, mais il peut également contraindre cette analyse. Pour cela, la grammaire doit tenir compte de ces traits dans sa description de la langue pour les exploiter au mieux de leurs possibilités.

Il faut d'abord savoir que la manipulation d'un trait par la grammaire est soumise à une déclaration préalable de ce trait dans un fichier lu par le moteur avant la grammaire proprement dite. Chacun des traits est composé d'un couple attribut-valeur, cette valeur pouvant bien entendu varier lors de la manipulation du trait. Dès lors, la déclaration d'un trait dans le fichier spécifique s'effectue par la détermination de l'attribut et de l'ensemble des valeurs [10] que cet attribut peut prendre :

```
attribut1:{valeur1,valeur2,...,valeurN},
attribut2:{valeur1,valeur2,...,valeurN},
...,
attributM:{valeur1,valeur2,...,valeurN}
```

On peut donc dire qu'un trait est défini par un attribut (`attribut1` pour le trait 1, `attribut2` pour le trait 2, `attributM` pour le trait M) dont les valeurs appartiennent strictement à l'ensemble `valeur1, valeur2, ..., valeurN`, dans lequel chaque valeur est nécessairement différente de toutes les autres. Dans les faits, l'application d'un trait à un **ancrage** déterminé ne peut spécifier qu'une seule valeur d'attribut. Lors de la déclaration d'un trait, il est possible de spécifier n'importe quelle valeur pour un attribut déterminé, que ce couple attribut-valeur ait une signification lexicale, morphologique, syntaxique, sémantique ou autre.

```
p1:{+,-},
p2:{+,-},
p3:{+,-}
```

FIG. 2.9 – Exemple de déclaration de traits dans une grammaire *XIP*.

L'exemple 2.9 montre la déclaration de trois traits d'attributs respectifs `p1`, `p2` et `p3`, dont la valeur peut être positive ou négative. Ce sont des traits morphologiques qui s'appliquent à des verbes ou à des pronoms personnels et qui en dénotent la personne. Ainsi, le pronom personnel *il* sera porteur d'un trait `p3 :+` tandis que la forme *dois* du verbe *devoir* dénotera les traits `p1 :+, p2 :+`.

[9] Pour éclaircir notre propos, nous appelons « ancrage » du trait l'entité à laquelle un trait est rattaché – que ce soit un lexème, un nœud lexical ou non lexical de l'arbre syntaxique partiel ou une dépendance syntaxique.

[10] On appelle « domaine de l'attribut » l'ensemble des valeurs que peut prendre l'attribut d'un trait.

Il existe toutefois un attribut qui fait exception à ce mode de déclaration et d'utilisation. En effet, il peut arriver que la déclaration de toutes les valeurs possibles pour un attribut ne soit pas envisageable. Dans ce cas, la déclaration du trait ne passe pas par un couple attribut-valeur. L'attribut est fixe (`$STACK`) et la valeur de cet attribut pourra être n'importe quelle chaîne de caractères. Le domaine de cet attribut n'est donc pas réellement déterminé. Cette particularité est un moyen efficace d'attacher à un point d'ancrage de trait des données dépendant de l'analyse lorsqu'il s'agit d'informations qu'on ne peut anticiper. Ce trait peut aussi s'appliquer plusieurs fois à un même ancrage et posséder une valeur différente – et donc être porteur de plusieurs informations – sans que cela pose problème. Nous verrons au cours de la phase d'enrichissement que l'assignation de synonymes à une unité lexicale exploite cette fonctionnalité (*cf.* section 5.4 page 170).

À côté des attributs entrant dans la composition des traits, il en existe d'autres, appelés « attributs généraux ». Un attribut général est un attribut commun à un ensemble de traits qu'il regroupe. Déclaré comme générique à ces traits, il n'a pas de valeur définie, sinon une caractéristique booléenne indiquant en cours d'analyse si son ancrage possède ou non un des traits dont cet attribut général est le générique. L'intérêt de l'attribut général s'affirme dans la manipulation des traits, et plus précisément dans les conditions d'application de règles *XIP* sur des traits.

L'utilisation des traits dans *XIP* est soumise à différents types de conditions. En effet, pour déterminer si un trait est présent ou non sur un point d'ancrage, il s'agit de vérifier la valeur liée à l'attribut de ce trait sur le point d'ancrage. La condition peut ainsi porter sur la valeur d'un attribut général qui englobe le trait visé ou sur la valeur du trait lui-même pour constater la présence de ce trait sur le point d'ancrage. Si la vérification est positive, on dira que cet ancrage dénote le trait dont la valeur est vérifiée. Dès lors, on peut spécifier la dénotation de l'ancrage en posant des conditions sur les traits qui lui sont ou non associés.

Dans l'exemple de déclaration de traits de *XIP* de la figure 2.10 page suivante, nous présentons un attribut général `pers` générique à trois traits. Ces traits sont d'attributs `p1`, `p2` et `p3` qui peuvent recevoir les valeurs `+` et `-`. Suivant les règles que nous avons énoncées ci-dessus, et en fonction de cette déclaration de traits, un point d'ancrage de traits peut dénoter les traits suivants :

`p1:+` ou `p1:-`
et/ou `p2:+` ou `p2:-`
et/ou `p3:+` ou `p3:-`

Il est possible de poser différentes conditions sur un trait au départ du point d'ancrage de ce trait. Quatre types de condition se dégagent qui peuvent être exploités par des règles *XIP* pour manipuler les traits en fonction de la

Déclaration formelle :

```
AttGen:[
     Att1:{+},
     Att2:{+}
       ],
Att3:{val1,val2}
```

Exemple :

```
pers:[
     p1:{+,-},
     p2:{+,-},
     p3:{+,-}
]
```

FIG. 2.10 – Déclaration d'un attribut général dans une grammaire *XIP*.

valeur de leur attribut :

1. une condition de type `[Att:Val]` est satisfaite lorsque l'ancrage présente une valeur `Val` pour l'attribut `Att` (ou : lorsque l'ancrage dénote le trait `Att:Val`) ;
2. une condition de type `[Att:~Val]` est satisfaite lorsque l'ancrage ne présente pas la valeur `Val` pour l'attribut `Att` ;
3. une condition de type `[Att]` est satisfaite lorsque l'ancrage présente une valeur `Val` quelconque pour l'attribut `Att` ou bien, si `Att` est un attribut général, lorsque l'ancrage dénote au moins un des traits dont `Att` est le générique ;
4. une condition de type `[Att:~]` est satisfaite lorsque l'ancrage ne présente aucune valeur `Val` pour l'attribut `Att` ou bien, si `Att` est un attribut général, lorsque l'ancrage ne dénote aucun des traits dont `Att` est le générique.

Il nous reste enfin à étudier comment réaliser l'assignation des traits à leur point d'ancrage. Il y a deux modes d'assignation principaux dont le fonctionnement est fonction de la nature de l'ancrage [11].

Le premier de ces modes d'assignation précède toute analyse effectuée par *XIP* et découle de l'analyse morphologique. Il ne peut donc porter que sur les unités lexicales, puisqu'il n'y a pas d'arbre syntaxique ni de dépendances à ce moment. Les traits assignés aux lexèmes consistent en réalité

[11] Un troisième mode d'assignation existe, la percolation ou unification des traits sur les nœuds entre partie droite et partie gauche de la règle. Il est rare et nous ne l'utilisons pas. Nous n'en traitons donc pas ici.

en une traduction des étiquettes fournies par l'analyseur morphologique. Remarquons que si l'analyseur morphologique fournit des informations d'une tout autre nature que le type d'analyse à laquelle il est dédié, elles peuvent être traduites sans distinction par la grammaire de *XIP* et assignées sous forme de traits aux lexèmes.

On trouvera un exemple de ce mode d'assignation à la figure 2.5 page 67. La catégorie grammaticale est l'étiquette du nœud terminal correspondant. Cette assignation sera effective si la désambiguïsation catégorielle aboutit. La plupart des autres informations morphologiques sont simplement traduites sous forme d'attributs (`sg`, `fem`, `p1`, `p2`) mais certains subissent un traitement plus important : `SubjP` ou `IndP` par exemple sont traduites chacune par deux traits, respectivement `subj` et `pre`, `ind` et `pre`.

Bien que ce mode d'assignation ne concerne que les unités lexicales du texte traité, l'application de règles de désambiguïsation catégorielle entraîne la sélection d'une seule analyse morphologique, et limite donc les traits affectés à chaque lexème. Toutefois, ces traits seront dès lors affectés au nœud lexical correspondant à chaque lexème dans l'arbre syntaxique partiel, les lexèmes ayant à ce moment reçu une étiquette catégorielle.

C'est au travers du formalisme de règles *XIP* que l'on applique la seconde méthode d'assignation de traits à un ancrage. Puisque les règles *XIP* produisent un résultat attaché à des nœuds ou à des dépendances, cette méthode, de par son mode de fonctionnement, ne permet pas l'assignation d'un trait à un lexème. Une règle *XIP* qui fait référence à un ancrage permet d'assigner un trait à cet ancrage ou de supprimer un trait déjà assigné sur cet ancrage. Pour effectuer cette assignation ou cette suppression, on utilise respectivement les formules `[Attribut=Valeur]` et `[Attribut=~]`. Notons qu'il est possible de combiner conditions sur les traits d'un ancrage et assignations de traits sur ce même ancrage. Pour illustrer ce mode d'assignation, nous prenons une règle issue de la partie de la grammaire consacrée à l'extraction des dépendances

```
if (
      VARG[indir,pron,left](#1,#2[dat,acc]) &
      ~VARG[dir](#1,#)
   )
VARG[indir=~,dir=+](#1,#2)
```

dont nous pouvons dire qu'elle permet de retirer (~) le trait *indirect* (`indir`) et d'assigner (`+`) le trait *direct* (`dir`) à une dépendance de type complément de verbe (`VARG`) si d'une part il existe entre un verbe quelconque (`#1`, les dépendances impliquant un verbe plaçant toujours le verbe en premier argu-

ment) et un argument dénotant les traits *datif* et *accusatif* (respectivement `dat` et `acc`) une relation de type complément indirect (`indir`) de verbe (`VARG`) impliquant un pronom (`pron`) placé à gauche du verbe (`left`), et d'autre part il n'y a pas de dépendance de type complément direct (`dir`) de verbe entre ce même verbe (`#1`) et un argument quelconque (`#`).

Situation d'application de la règle ci-dessus :
Julien **se** détache très tôt du christianisme.
`VARG[DIR](détache,se)`

Situation où la règle ci-dessus ne s'applique pas :
Antoine **se** donna *la mort* dans Alexandrie assiégée.
`VARG[INDIR](donner,se)`
`VARG[DIR](donner,mort)`

FIG. 2.11 – Utilisation d'une règle de construction de dépendance.

Nous avons indiqué les différentes qualités qui nous amènent à choisir cet analyseur syntaxique : sa robustesse permettant de traiter toutes les phrases de nos corpus, sa souplesse qui autorise des modifications, corrections et enrichissements de la grammaire, son mode de gestion des traits et dépendances qui admet l'incursion de la sémantique dans son analyse. De plus, le formalisme de *XIP* nous permet un stockage ordonné et indexé de tous les résultats obtenus, sous forme de nœuds, de traits et de dépendances qui représentent la structure du texte analysé, structure dont la profondeur est fonction du type d'analyse demandé. En particulier, la possibilité d'associer des traits littéraux aux dépendances (`$STACK`) nous ouvre certaines perspectives d'association entre un nœud et une **expression synonymique** non négligeables[12]. Cet outil nous semble donc adéquat pour le type d'application auquel nous désirons l'assigner.

2.3 Désambiguïsation sémantique lexicale

2.3.1 Aperçu des précédentes méthodes

Introduction

Une application qui entend gérer le sens du texte dans une base documentaire à travers l'analyse linguistique des énoncés qui la composent ne peut se passer d'une procédure visant à identifier la signification des

[12] En effet, il suffit d'assigner au nœud considéré un trait littéral d'attribut `$STACK` dont la valeur correspond à la chaîne de caractères de l'expression synonymique.

unités de sens manipulées dans les documents de cette base. La désambiguïsation sémantique (*Word Sense Disambiguation,* ***WSD***), qui permet de décider du sens des unités lexicales dans un texte, constitue une « tâche intermédiaire »[Wilks et Stevenson, 1996] essentielle dans le cadre de nombreux processus de traitement automatique de la langue [13], et principalement dans les applications visant la compréhension de texte en langage naturel [Ide et Véronis, 1998].

Du fait de cette grande variété d'intérêts, les difficultés liées à la problématique de la désambiguïsation sémantique ont très tôt été identifiées. Toutefois, les solutions qui ont été proposées dans chaque domaine ont également été multiples et très diverses, en fonction des besoins et des savoirs afférents à chacune des matières concernées. La définition du problème elle-même ne fait pas l'unanimité. En effet, si un consensus est atteint pour définir la désambiguïsation sémantique comme l'association d'un mot apparaissant dans un contexte avec sa signification ou sa définition – laquelle peut être distinguée des autres définitions qu'on peut attribuer à ce mot –, en revanche le même accord n'existe pas pour ses sous-tâches.

Il s'agit en effet de déterminer d'abord l'ensemble des sens que peut prendre chaque mot dans la langue. [Kelly et Stone, 1975] montre que l'attribution objective d'un sens particulier à une unité lexicale polysémique dans un contexte donné n'est pas chose aisée. Toutefois, à l'heure actuelle, les travaux en désambiguïsation sémantique s'effectuent principalement à partir de sens prédéfinis, grâce à diverses ressources lexicales et sémantiques. D'autre part, l'assignation d'un sens particulier à une unité lexicale exploite deux informations principales, à savoir le contexte d'apparition des occurrences de chaque mot, et une ou plusieurs bases de connaissances externes qui permettent de mettre en rapport les mots en contexte avec leur sens. C'est sur la nature de la base de connaissance que survient ici le désaccord, certaines méthodes privilégiant des ressources d'un ordre plutôt lexical pour fournir ces données (*knowledge-driven word sense disambiguation*), d'autres leur préférant des informations sur le contexte provenant de corpus aux unités lexicales préalablement désambiguïsées (*corpus-based word sense disambiguation*).

Parmi les différentes méthodes de détermination de la signification des mots en contexte, nous ne nous intéressons toutefois qu'à celles qui correspondent aux restrictions que nous nous sommes fixées dans le cadre de cette thèse, à savoir les méthodes qui se fondent sur des critères linguistiques pour effectuer la tâche qui leur est confiée. Par ailleurs, la désambiguïsation sémantique s'inscrit ici dans le contexte d'un processus d'enrichissement du texte qui lui est soumis, et doit de ce fait permettre la sélection de l'information

[13] L'état de l'art de [Ide et Véronis, 1998] recense six grands domaines pour lesquels la compréhension du langage est un enjeu intermédiaire : traduction automatique, recherche d'information et navigation hypertexte, analyse thématique et du contenu, analyse grammaticale, traitement de la parole, traitement du texte.

lexico-syntaxique la plus riche et la plus précise, ce qui implique l'utilisation d'une ressource lexicale bien structurée. De plus, le texte qui est soumis à la désambiguïsation sémantique est libre et susceptible d'atteindre un volume important, d'où une nécessité de robustesse. Ces exigences limitent donc l'horizon des systèmes de désambiguïsation sémantique auxquels nous nous intéressons. Nous ne ferons qu'évoquer succinctement les autres.

Un précurseur : la traduction automatique

Dans les années qui ont suivi la seconde guerre mondiale, la traduction automatique fut la première spécialité à s'intéresser aux problèmes liés à la polysémie des mots. Très vite, les travaux qui s'y consacrèrent admirent l'importance déterminante du contexte d'un mot à désambiguïser (que nous appelons « **cible** », *target*) [Weaver, 1949, Kaplan, 1955], et ensuite l'influence très marquée des relations syntaxiques entre la cible et son contexte [Reifler, 1955].

Par la suite, les besoins de connaissance d'un univers plus large pour effectuer les distinctions de sens ont initié deux tendances : tout d'abord la réduction du traitement à des domaines restreints, ce qui amène l'utilisation d'un lexique spécialisé dont la polysémie est limitée et donc la désambiguïsation facilitée [Panov, 1960], mais cette approche n'est pas envisageable dans le cadre du texte tout-venant ; ensuite, inspirée par la notion de langue-pivot, l'idée développée par [Masterman, 1957, Masterman, 1961] d'une abstraction de la forme de surface en concepts dans un réseau sémantique structuré, qui permet de choisir le sens correspondant au concept le plus proche du contexte. Cette seconde tendance très novatrice préfigure le travail de l'intelligence artificielle en désambiguïsation sémantique. Par ailleurs, la direction prise entre autres par les approches décrites dans [Pimsleur, 1957] et dans [Madhu et Lytle, 1965], qui exploitent l'étude quantitative de la polysémie du lexique ainsi que la probabilité d'apparition d'un sens dans un contexte donné pour effectuer le choix du sens, inaugure l'application de méthodes statistiques au domaine.

La veine de l'intelligence artificielle

La plupart des méthodes de sélection du sens en intelligence artificielle n'ont donné lieu qu'à des implémentations extrêmement limitées au niveau du vocabulaire et au niveau du contexte. Cette limitation ne permet pas d'appliquer ces méthodes à du texte réel. Toutefois, certaines approches sont intéressantes par leur principe, qui pourra être réutilisé dans d'autres perspectives.

Le réseau sémantique de [Masterman, 1961] permet d'abstraire le sens des phrases dans une langue-pivot composée de concepts fondamentaux. Autour d'une centaine de types de concepts primitifs (THING, DO...), un dictionnaire de 15 000 concepts est construit sous la forme d'un réseau hiérarchique qui autorise l'héritage vertical descendant des propriétés. Le choix des sens est implicite et s'effectue au niveau de la phrase : ce sont les nœuds du réseau correspondant aux concepts les plus proches qui sont activés, fournissant ainsi la signification de chacune des unités lexicales. Les approches symboliques ultérieures qui visent l'exploitation d'un réseau sémantique vont s'atteler à donner une étiquette sémantique aux liens qui constituent le réseau [Quillian, 1968], ainsi qu'à fournir un cadre informationnel sur les unités lexicales et leurs relations entre elles [Hayes, 1977], mais conservent le principe du chemin le plus court entre deux nœuds comme meilleur choix de sens.

Le système proposé dans [Hirst, 1987] exploite lui aussi un réseau sémantique et des cadres informationnels liés aux unités lexicales afin de définir le chemin le plus court entre deux nœuds, mais il introduit en plus un mécanisme appelé « mots polaroïds » (*polaroid words*) qui élimine progressivement les sens qui ne peuvent être appliqués à cause d'indices fournis soit par une analyse syntaxique, soit par l'information présente dans le cadre informationnel. Il note toutefois que si la phrase sort du cadre informationnel défini, aucune décision ne pourra être prise.

Tout en abandonnant le principe du réseau sémantique, [Wilks, 1975] insiste lui aussi sur les relations que la cible entretient avec son entourage contextuel. Pour chaque unité lexicale, il établit un réseau de préférences sémantiques sous la forme de restrictions de sélection régissant la combinaison syntaxique et sémantique de la cible avec d'autres lexèmes. Ces restrictions peuvent progressivement être assouplies dans les cas où les règles les plus strictes n'aboutissent pas à un résultat.

Pour la désambiguïsation sémantique de sa méthode de compréhension du langage naturel, [Dahlgren, 1988] utilise plusieurs informations, dont des syntagmes figés, des restrictions de sélection syntaxico-sémantiques et un moteur de raisonnement « de bon sens », qui consiste à chercher un ancêtre commun à deux mots appartenant au contexte dans une ontologie, comme [Resnik, 1995] le fera également. Dahlgren note que la moitié des désambiguïsations sont effectuées par ce module ontologique, que les restrictions de sélection des verbes sont une importante source d'information pour la désambiguïsation des noms.

Suite à la notion d'« **amorçage sémantique** »[14] (*semantic priming*), le courant connexionniste va exploiter les réseaux sémantiques selon des modèles de « propagation d'activation » (*spreading activation*), c'est-à-dire que dans un réseau sémantique, les concepts sont activés lorsqu'ils sont mentionnés dans le document, et cette activation est transmise aux nœuds qui sont connectés à ces concepts. L'activation se délite progressivement, mais il est possible qu'un même nœud soit activé par différentes sources, ce qui renforce son activation par rapport aux autres. Bien que ces approches pondérées ne correspondent pas à une méthode linguistique, [Bookman, 1987] a introduit dans le réseau des traits sémantiques (opposition fondamentales, durée, lieux...) pour permettre de contraindre plus précisément la sémantique des nœuds activés. Ces approches n'ont cependant pas été menées à une échelle suffisante pour être exploitables dans une application en taille réelle.

Les méthodes basées sur des ressources lexicales

Les méthodes de désambiguïsation sémantique avancées dans le domaine de l'intelligence artificielle présentent surtout le défaut d'une couverture lexicale insuffisante. Dès que les possibilités matérielles ont permis la gestion de grands volumes de données, les recherches en désambiguïsation sémantique se sont attachées à utiliser des ressources lexicales de grandes dimensions. [Michiels, 1982] attire notamment l'attention sur la richesse de l'information contenue dans ces ressources. Il insiste sur l'intérêt que ces données représentent pour le traitement du langage en général, et pour le traitement de la sémantique en particulier.

Les premières tentatives ont été faites avec les **dictionnaires au format électronique**, dont on essayait d'extraire une information lexicale et sémantique. Cependant, une information rigoureuse n'est pas facile à obtenir, ces dictionnaires présentant deux défauts majeurs : ils comportent de grandes incohérences [Kilgarriff, 1994] et ils sont conçus pour être utilisés par des humains, sans tenir compte des besoins logiciels. Dès lors, les approches appliquent un principe de sécurité, préférant donc la robustesse à la finesse. L'idée force de ce principe est qu'un mot polysémique voisin d'un autre mot dans un contexte possède celui de ses sens qui se rapproche le plus du sens de son voisin. Les indices de proximité entre les sens de deux mots varient en fonction des méthodes. Ce principe favorise bien entendu les modèles statistiques, même si des notions plus linguistiques peuvent y être adjointes dans certaines approches.

[14] L'amorçage sémantique (*semantic priming*) est une théorie psycholinguistique selon laquelle l'introduction d'un concept dans un énoncé va influencer et faciliter la compréhension de concepts ultérieurs sémantiquement reliés [Meyer et Schvaneveldt, 1975].

[Lesk, 1996] imagine un système qui génère une base de connaissances à partir d'un dictionnaire de langue, constituant pour chaque sens de chaque lexème une « signature » composée de la liste des mots apparaissant dans la définition de ce sens. La désambiguïsation de la cible se fait par sélection du sens qui présente la plus grande intersection avec les signatures des mots du contexte. [Wilks et al., 1993] améliore cette méthode fruste en augmentant la part accordée aux statistiques : il calcule la fréquence de co-occurrence des mots dans les définitions afin de définir un degré de relation entre les mots. [Véronis et Ide, 1990] reprend aussi la méthode de Lesk et l'exploite dans un réseau neuronal où chaque mot est relié à ses sens, qui sont reliés à chaque mot de leur définition, eux-même reliés à chacun de leurs sens, etc.

[Cowie et al., 1992] s'intéresse à une information supplémentaire, à savoir les catégories sémantiques définies dans le *Longman Dictionary of Contemporary English* (***LDOCE***) qui sont de deux types : les *box codes* qui présentent des catégories sémantiques (abstrait, humain...) et les *subject codes* qui correspondent à des domaines d'application (économie, ingénierie...). Ils améliorent la méthode de Lesk en imposant au sens sélectionné une correspondance de trait sémantique avec son contexte. Il reste que cette information sémantique n'est pas systématique dans le LDOCE. Plus grave, le LDOCE, comme la plupart des dictionnaires électroniques, manque cruellement d'informations pragmatiques permettant d'établir des liens entre les unités lexicales et entre les informations dont elles sont porteuses.

Les **thesaurus** sont le deuxième type de ressources lexicales, plus systématiques que les dictionnaires et fournissant des relations essentiellement synonymiques entre les mots. Chaque occurrence d'un mot dans une catégorie d'un thesaurus correspond à un de ses sens, chaque catégorie rassemblant des mots ayant approximativement le même sens. Cette particularité de conception a valu aux thesaurus d'être très tôt exploités pour le traitement automatique de la sémantique, notamment pour la constitution du réseau sémantique de [Masterman, 1957] (voir 2.3.1 page 77).

Les méthodes qui exploitent les thesaurus sont généralement axées sur une information statistique importante, à l'image de [Yarowsky, 1992], qui établit un modèle statistique basé sur le contexte. Chaque catégorie du *Roget's Thesaurus* est considérée comme une classe de mots. À partir de chacun des éléments de chaque classe, Yarowsky construit un ensemble contextuel de cent mots extraits d'un corpus et établit la probabilité statistique que chaque mot de la classe et chacun des cent mots de son contexte soient co-occurrents. La désambiguïsation sémantique est effectuée par l'application de la formule de Bayes sur la probabilité pour chaque classe contenant la cible d'être choisie.

Enfin, les **dictionnaires informatiques**, exploitables seulement par une application logicielle, rassemblent sous la forme de bases de connaissances

des informations plus ou moins liées au lexique [15] au niveau morphologique, syntaxique et/ou sémantique. La désambiguïsation sémantique exploite essentiellement *WordNet* [Fellbaum, 1998b], dont l'information peut se rapprocher tantôt d'un dictionnaire (définitions), tantôt d'un thesaurus (groupes de mots quasi-synonymes appelés *synsets*, hiérarchie conceptuelle), ou bien d'un réseau sémantique (relations **hyponymiques**, **méronymiques**, **antonymiques**), etc. On notera toutefois que cette ressource ne contient pas d'information syntaxique.

[Voorhees, 1993] exploite l'information hyponymique de *WordNet* dans une perspective de recherche d'information en cherchant, grâce à la construction des sous-graphes hyponymiques de chaque mot, à établir des similitudes sémantiques entre les mots de la requête et ceux de sa réponse. Ces similitudes sont obtenues grâce au décompte des *synsets* des unités composant la requête et ceux des documents. Cependant, aucun choix réel de sens fin n'est effectué par cette méthode, seulement un rapprochement de deux mots. [Sussna, 1993], dans une semblable perspective de recherche d'information, attribue un poids à chaque type de relation entre deux unités lexicales et donne à chaque lexème une mesure liée au nombre de relations de même type qui la relient à d'autres. Ces mesures servent de base à un calcul appliqué aux chemins qui relient deux unités lexicales voisines dans un texte, et le sens choisi est celui qui obtient le meilleur résultat. Sussna observe l'importance de sens proches dans un même contexte. Il note également l'intérêt d'utiliser d'autres relations sémantiques que le classique IS-A. S'appuyant sur les travaux décrits dans [Dahlgren, 1988] en intelligence artificielle, [Resnik, 1995] recherche dans la hiérarchie IS-A un terme générique commun à deux lexèmes (ou plus) d'un texte et calcule la longueur du chemin permettant de déterminer la portion d'information commune aux lexèmes. Toutefois, il se distingue de [Sussna, 1993] en considérant que la distance entre deux nœuds du réseau varie selon le type de relation qui les unit.

Nous notons que ces différentes méthodes ne s'appliquent qu'aux substantifs, et que la distinction entre les sens est effectuée par le calcul d'une distance ou d'un poids pour chaque sens de chaque mot, qui permettent de rapprocher ou d'opposer des données sémantiques. Il est toutefois intéressant de constater le bénéfice apporté d'une part par les différentes relations sémantiques qui constituent le réseau, et d'autre part par la distinction de l'importance qu'il faut apporter à ces relations sémantiques. Cependant, le contenu de *WordNet* n'est pas parfait, la distinction des sens elle-même étant

[15] Il s'agit de ressources développées à la main dans un format adapté non à l'usage humain, mais à une exploitation par ordinateur. Certaines ressources ne sont pas attachées à une langue particulière : *CyC*, *Mikrokosmos*. Pour l'anglais, il y a principalement *ACQUILEX* [Briscoe, 1991], *COMLEX* et *WordNet* [Miller et al., 1990, Fellbaum, 1998b]. Pour le français *AlethDic* [GENELEX, 1994] et *EuroWordNet* [Vossen, 1998] (dans sa partie francophone, voir 3.4.3 page 120).

souvent trop fine[16], et l'information syntaxique manquant cruellement.

Un autre type de ressources lexicales informatiques existe, qui ne décrit pas les différents sens des mots de manière énumérative, mais sous la forme de règles qui décrivent les sens de manière relative. C'est le **lexique génératif** [Pustejovsky, 1991]. Divers travaux tendent à utiliser ce type de ressources génératives pour effectuer un travail de désambiguïsation sémantique [Viegas et Bouillon, 1994, Viegas et al., 1999]. Toutefois, l'absence de dictionnaire génératif pour le français et le manque d'information permettant d'aboutir à un enrichissement de texte nous ont amené à écarter ces méthodes.

Les méthodes basées sur l'analyse de corpus

Les méthodes basées sur l'étude de grands corpus textuels s'adaptent bien à l'élaboration de modèles statistiques qui reposent sur l'étude de fréquences rencontrées dans les textes. Cependant, des méthodes linguistiques basées sur des observations et sur la construction de règles à partir de ces observations ont abondamment utilisé les corpus pour obtenir l'information dont elles avaient besoin. [Weiss, 1973] a démontré sur cinq mots et un corpus d'une vingtaine de phrases pour chaque mot que des règles de désambiguïsation sémantique pouvaient être extraites de phrases étiquetées sémantiquement. [Kelly et Stone, 1975] a suivi son exemple : à partir d'un corpus de 500 000 mots, Kelly et Stone ont extrait manuellement des règles de désambiguïsation sémantique pour chaque sens de 1800 mots polysémiques. Ces règles exploitaient des indices tels que la **collocation**, les relations syntaxiques et l'appartenance à une même catégorie sémantique. Bien que réalisés sur une petite échelle, ces tests donnaient d'excellents résultats.

Cependant, les modèles statistiques se sont rapidement imposés lorsque le volume de données contenues dans les corpus a commencé à devenir réellement important : [Black, 1988], par exemple, a extrait des arbres de décision sémantique d'un corpus de 22 millions de mots dont il avait étiqueté environ 2 000 occurrences de cinq lexèmes. Toutefois, cette méthode elle-même, comme celles de Kelly et Stone, met en évidence les difficultés d'exploiter des corpus pour un traitement sémantique. En effet, il s'agit non seulement d'étiqueter manuellement ces textes, mais aussi d'obtenir des documents qui comportent des occurrences de chacun des sens de chacun des mots du lexique, et cela en nombre suffisant pour pouvoir inférer des normes de comportement, que ce soit dans les méthodes linguistiques ou statistiques. Les tentatives d'améliorations ont donc porté sur deux problèmes.

[16] La distinction des sens est parfois trop subtile, et n'est pas forcément évidente même pour un utilisateur humain.

Il a d'abord fallu trouver des moyens d'étiqueter par le sens ces grandes bases textuelles par des techniques automatiques. Une solution proposée est l'**amorçage** (*bootstrapping*), qui comporte une phase d'apprentissage d'informations qui permettront ultérieurement un étiquetage sémantique automatique. La technique de l'amorçage est toutefois généralement soumise à des données quantitatives, que ce soit la méthode de [Hearst, 1991], qui propose d'extraire des données statistiques du contexte des dix à trente premières occurrences – manuellement étiquetées – de chaque mot dans le corpus, ou celle de [Schütze, 1992], basée sur des tétragrammes et sur la représentation vectorielle du sens de l'ensemble des mots formant le contexte de la cible.

Une autre méthode a été testée pour éviter l'étiquetage manuel des corpus : l'utilisation de corpus bilingues alignés, qui permettent de distinguer les différents sens d'un mot par leur différentes traductions [Gale et al., 1993, Dagan et al., 1991]. Cependant, les sens possédant une même traduction ne peuvent être distingués. De plus, les corpus bilingues alignés correspondent le plus souvent à un domaine et à un registre de langue particuliers, et ne se révèlent pas forcément très représentatifs de la langue.

L'autre obstacle qu'il a fallu franchir concerne le manque de données représentatives de l'ensemble des sens de l'ensemble des mots du lexique. Le **lissage** (*smoothing*) est une méthode utilisée pour éviter que la probabilité d'apparition d'un sens rare et non représenté dans le corpus soit égale à zéro. Elle s'appuie sur des fréquences de co-occurrences et à ce titre ne retient pas notre attention. Les modèles basés sur des classes de mots (*class-based models*) s'appuient sur l'hypothèse que des unités lexicales appartenant à une même classe de mots peuvent être utilisés indifféremment. Ce type d'approche est tenté à partir de classes de mots appartenant à la taxinomie de *WordNet* [Resnik, 1992]. Aux catégories du *Roget's Thesaurus* [Roget et Dutch, 1972, Yarowsky, 1992] ou aux *codes* du *LDOCE* [Slator, 1992, Liddy et Paik, 1992]. Cependant, cette hypothèse trop stricte est à l'origine de la perte de toute information qui n'est pas partagée par l'ensemble des membres de la classe. Un autre modèle s'est donc développé, basé sur les similarités (*similarity-based methods*), qui part du même postulat, mais ne construit pas de classes fixes, chacun des lexèmes appartenant potentiellement à plusieurs ensembles de mots similaires [Dagan et al., 1993].

2.3.2 Problèmes et solutions

Au cours de cet examen des approches utilisées pour gérer le problème de l'ambiguïté sémantique lexicale, nous avons fait plusieurs observations qui doivent nous servir pour le choix d'une méthodologie applicable à nos besoins.

Tout d'abord, il a été remarqué très tôt que l'étude du contexte de la cible constituait la principale information qui permet d'en sélectionner le sens adéquat. Par la suite, on a constaté l'importance prépondérante des éléments du contexte syntaxiquement liés à la cible pour effectuer ce choix.

Par ailleurs, nous voulons utiliser une méthode de désambiguïsation sémantique dans une perspective d'enrichissement et d'expansion de texte, c'est-à-dire que le choix d'un sens correct doit également permettre la sélection d'une information qui s'y rattache. Il s'agit donc d'utiliser une ressource lexicale descriptive qui comporte un maximum de données morphologiques, syntaxiques et bien entendu sémantiques. Seul un dictionnaire électronique est actuellement capable de fournir ce type d'information, les ressources de type *WordNet* ne présentant aucune information syntaxique qui permettrait de gérer le contexte syntaxique de la cible. Toutefois, il n'est pas question de rejeter certaines utilisations des données sémantiques que contiennent cette catégorie de ressources, et notamment les relations qu'elles sont capables de décrire entre les mots.

Enfin, l'importance des informations linguistiques contenues dans les corpus étiquetés n'est pas négligeable, comme l'ont montré [Weiss, 1973, Kelly et Stone, 1975]. L'exploitation de tels corpus, pour peu qu'ils existent, ne doit en aucun cas être délaissée.

Pour ces différentes raisons, nous avons décidé de nous pencher particulièrement sur la méthodologie de désambiguïsation sémantique développée à XRCE. Cette méthode, simple à l'origine, exploitait à ses débuts une matrice HMM dans une fenêtre contextuelle de un mot avant et après la cible [Segond et al., 1997]. Une collaboration avec le CELI [Dini et al., 1998] a permis une amélioration du système grâce à l'utilisation de l'algorithme d'apprentissage de Brill, et l'extension de la fenêtre contextuelle aux unités lexicales reliées syntaxiquement à la cible. Enfin, la maturité qui nous intéresse a été atteinte [Segond et al., 1998] : le système actuel présente l'avantage d'exploiter l'information contextuelle lexicale et syntaxique, qui utilise un dictionnaire électronique constitué de manière à éviter les incohérence habituelles de ses homologues et présentant de nombreux exemples répartis par sens, exploitables comme un corpus étiqueté. Enfin, cette méthode utilise dans la création de ses règles l'information taxinomique de *WordNet* pour l'anglais, celle d'*AlethDic* pour le français afin d'apporter aux règles de désambiguïsation lexicales l'information sémantique qui leur manque et mettre les lexèmes en relation les uns avec les autres.

2.3.3 La méthode développée par CELI et XRCE

La méthode déjà expérimentée au CELI et à XRCE consiste à effectuer une analyse linguistique – essentiellement lexicale et syntaxique – de l'en-

vironnement typique d'un mot polysémique dans chacun de ses sens pour constituer un ensemble de règles de discrimination de ces sens. On comprendra dès lors que l'efficacité de cette méthode est liée à la qualité des informations contextuelles présentes dans le dictionnaire, ce qui justifie l'attention avec laquelle le choix de cette ressource est réalisé.

La méthode que nous décrivons ici a d'abord été implémentée et validée dans le cadre des compétitions SENSEVAL [Dini et al., 2000] et ROMANSEVAL [Segond et al., 2000] (*cf.* section 2.3.4 page 96) puis perfectionnée suite aux résultats obtenus [Brun, 2000]. Les outils d'analyses en sont *IFSP* (*Incremental Finite-State Parser*) pour l'analyse syntaxique (*cf.* section 2.2.2 page 62) et le *Xerox Morphological Analyser*, transducteur morpho-lexical contenant l'information morphologique correspondant au lexique encodé pour l'analyse morpho-lexicale. La ressource lexicale est la version électronique d'un dictionnaire bilingue français-anglais publié conjointement par les éditions Hachette et Oxford University Press [Corréard et Grundy, 1994] [17].

Remarquons certaines caractéristiques de ce module de désambiguïsation sémantique. Tout d'abord, il s'agit d'un système basé sur une ressource lexicale. Le dictionnaire utilisé comme référence est un dictionnaire bilingue, très riche en exemples et en indications contextuelles. Bien que cette particularité soit très utile dans des application multilingues, et notamment l'aide à la traduction, un tel dictionnaire s'applique peu à nos besoins (identification de synonymes et autres relations sémantiques, dérivations, catégorisation...). Il nous faut donc choisir une autre ressource lexicale de référence qui se prête mieux à nos besoins (*cf.* chapitre 3 page 99). Ensuite, l'intégration de ce prototype s'est faite au sein d'une plate-forme développée à XRCE. Cette plate-forme regroupe différents outils d'analyse textuelle sous le nom de *XeLDA* (*Xerox Linguistics Development Architecture*). Or les outils et ressources de *XeLDA* ne présentent ni la souplesse, ni les fonctionnalités, ni toutes les qualités des outils dont nous avons parlé précédemment. Il nous faudra donc nous extraire de cette architecture et présenter une nouvelle application basée sur les principes qui nous intéressent dans cette méthode. Enfin, deux étapes sont nécessaires à cette méthode de désambiguïsation sémantique : l'extraction des règles de désambiguïsation, c'est-à-dire la génération, à partir d'informations issues de l'analyse du dictionnaire, des règles dans un formalisme rigoureux compatible avec le système de désambiguïsation sémantique applicatif, et l'application de ces règles au cours de l'analyse textuelle.

[17] Nous désignerons désormais le dictionnaire *Oxford Hachette French Dictionary* par l'acronyme *OHFD*.

La méthode d'extraction des règles

Le principe de la méthode de désambiguïsation qui nous intéresse consiste à analyser les données syntaxico-sémantiques de chaque sens de chaque lemme pour relever les indices permettant d'effectuer une discrimination de sens lorsque ces indices sont détectés dans le contexte de l'unité lexicale à désambiguïser. De la sorte, chaque règle de désambiguïsation est attachée à un seul sens et elle associe ce sens à une information précise considérée comme typique de cette acception du lemme par le dictionnaire. Nous illustrons l'extraction des règles de désambiguïsation sémantiques au travers de l'entrée de *coucher* dans l'*OHFD* (*cf.* figure 2.12).

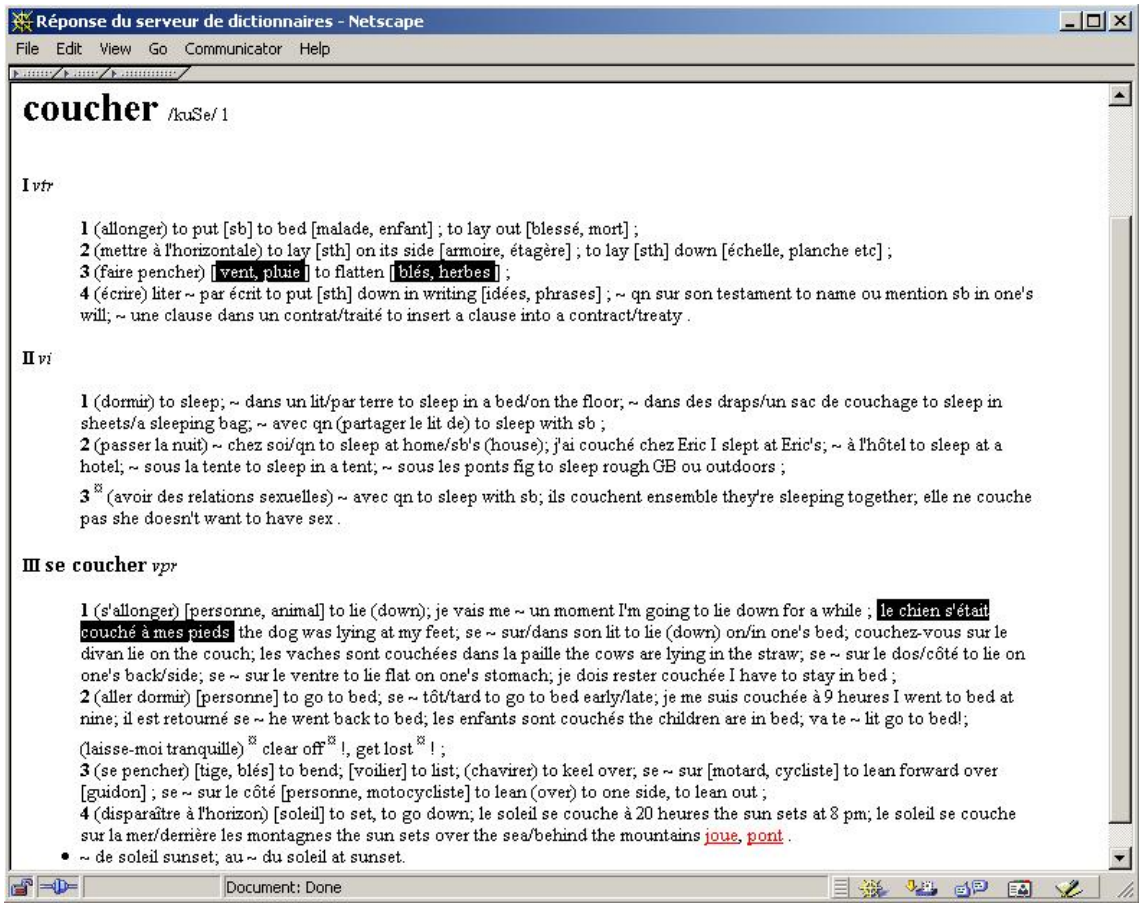

FIG. 2.12 – Entrée *coucher* dans la version électronique du dictionnaire *OHFD*

Un examen des données linguistiques présentes dans le dictionnaire utilisé constitue donc le principe essentiel de la méthode de constitution d'un ensemble de règles lexicales de désambiguïsation. Cet examen, qui peut va-

rier selon la ressource utilisée, se penche sur diverses informations. La base des exemples constitue la principale source de données, comme le montre la phrase *Le chien s'était couché à mes pieds* (sens *III.1* dans notre cas illustratif). En effet, on peut voir que l'analyse de cet exemple permet de dégager certaines relations syntaxiques typiques de cette acception du lemme *coucher*. On en extrait une règle impliquant cette acception *III.1* si dans le contexte du lemme *coucher* on trouve les mots *chien* ou *pied*, *chien* entretenant avec le lemme *coucher* une relation de type sujet d'un verbe *coucher* en construction pronominale (`SUBJREFLEX(chien,coucher)`) et *pied* étant le complément prépositionnel modifieur de verbe introduit par la préposition *à* (`VMODOBJ(coucher, à, pied)`).

De telles règles peuvent être construites pour chaque sens de chaque entrée du dictionnaire qui comporte un exemple, par simple analyse syntaxique de cet exemple. Chaque dépendance mettant en cause le **mot-vedette** constitue un élément de la condition d'application de la règle pour le sens dans lequel l'exemple apparaît.

```
[##]> echo "Le chien s'était couché à mes pieds." |
        ifsp -l french

SUBJREFLEX(chien,coucher)
VMODOBJ(coucher,à,pied)
  Le chien s' était couché à mes pieds .
  [SC  [NP Le chien NP]/SUBJ :v s' était couché SC]
  [PP à mes pieds PP] .
```

FIG. 2.13 – Analyse syntaxique d'un exemple à l'aide de *IFSP*.

Dans le même ordre d'idée, le dictionnaire *OHFD* présente régulièrement des indicateurs de collocation pour un sens donné dans une entrée. Une collocation décrit l'association habituelle d'une unité lexicale avec une autre dans un énoncé. Lorsque la collocation du mot-vedette avec un autre lexème est jugée suffisamment discriminante d'une acception par le lexicographe chargé de l'entrée, un indicateur de collocation est utilisé dans le dictionnaire. L'*OHFD* comporte de plus cette particularité qui veut que la position de l'indicateur de collocation par rapport à la traduction de la vedette soit indicative de la construction syntaxique habituellement entretenue entre ce mot-vedette et sa collocation. Ces informations sont d'une précieuse aide dans la constitution de règles de désambiguïsation sémantique. Notre cas illustratif nous montre que dans le cadre du sens *I.3* du lemme *coucher*, la collocation *vent* (ou *pluie*) est un sujet typique tandis que *blés* (ou *herbes*) est un objet distinctif de *coucher* pour ce même sens. On peut donc obtenir deux règles de désambiguïsation impliquant le sens *I.3* si le contexte du

lemme *coucher* offre soit le mot *vent* (ou *pluie*) comme sujet de coucher (`SUBJ(vent,coucher)`), soit le mot *blés* (ou *herbes*) comme objet de *coucher* (`DOBJ(coucher,blés)`). Cette démarche permet d'obtenir les dépendances syntaxiques qui correspondent à l'analyse de phrases générées à partir de ces informations de collocation (*cf.* exemple 2.14). Toutefois, le fait de rester au niveau syntaxique et de ne pas construire réellement les phrases permet d'éviter les problèmes liés à la génération (flexions et règles orthographiques).

Dépendances générées à partir des collocations :

```
SUBJ(vent,coucher)
SUBJ(pluie,coucher)
DOBJ(coucher,blés)
DOBJ(coucher,herbes)
```

Dépendances extraites des phrases correspondantes par l'analyser syntaxique *IFSP* :

```
[##]> echo "Le vent couche les blés. La pluie
couche les herbes." | ifsp -l french

SUBJ(vent,coucher)
DOBJ(coucher,blé)
      Le vent couche les blés .
SUBJ(pluie,coucher)
DOBJ(coucher,herbe)
      La pluie couche les herbes .
```

FIG. 2.14 – Dépendances obtenues d'après les collocations.

De telles règles sont lexicales, c'est-à-dire qu'elles s'appliquent exclusivement à la désambiguïsation sémantique du lexème pour lequel elles ont été construites.

Le travail que nous avons effectué en désambiguïsation sémantique lexicale nous a d'abord amené à étudier les possibilités d'amélioration de la méthode que nous avons exposée. Ces améliorations se sont faites selon trois perspectives. La première est la recherche de la généralisation des règles afin de sortir du champ restreint du lexique strict et limité des exemples et collocations. En cela, nous avons pris exemple sur le travail effectué par [Brun, 2000] dans la version anglaise du prototype. Là où le vocabulaire était remplacé par une ou plusieurs des classes sémantiques de *WordNet* [18] à l'intérieur des règles, nous avons fait de même à partir des classes taxinomiques

[18] *WordNet* est une ressource lexico-sémantique anglophone [Fellbaum, 1998b, Voorhees, 1993] dont deux particularités ont retenu notre attention : la mise en relation des unités lexicales en fonction de leur sens et selon des relations de type synonymique, hy-

de *AlethDic* [19]. Si nous reprenons l'exemple de notre cas illustratif (*Le chien s'était couché à mes pieds*), la règle généralisée correspondante serait : l'acception *III.1* du lemme *coucher* est choisie si dans le contexte de *coucher* on trouve un mot qui appartient à la classe *ANIMAL* ou à la classe *INSTRUMENT* (classes du mot *chien*) entretenant une relation `SUBJREFLEX` (sujet d'un verbe pronominal) avec *coucher* ou un mot de classes *FORME*, *MESURE*, *MEUBLE*, *CORPS* ou *VEGETAL* (classes du mot *pied*) entretenant une relation `VMODOBJ` (complément prépositionnel modifieur de verbe) avec *coucher*.

Règles lexicales pour l'exemple *Le chien s'était couché à mes pieds* :

```
coucher : SUBJREFLEX(chien,coucher)
           ==> #coucher.3.1#
coucher : VMODOBJ(coucher,à,pied)
           ==> #coucher.3.1#
```

Classes sémantiques :

chien	ANIMAL/INSTRUMENT
pied	FORME/MESURE/MEUBLE/CORPS/VEGETAL

Règles sémantiques pour l'exemple *Le chien s'était couché à mes pieds* :

```
coucher : SUBJREFLEX(ANIMAL/INSTRUMENT,coucher)
           ==> #coucher.3.1#
coucher : VMODOBJ(coucher,à,FORME/MESURE/MEUBLE/CORPS/VEGETAL)
           ==> #coucher.3.1#
```

FIG. 2.15 – Généralisation des règles de désambiguïsation sémantique.

La deuxième perspective d'amélioration est centrée sur l'exploitation d'autres informations présentes dans le dictionnaire. Les possibilités de recherche dans cette voie sont bien entendu limitées par deux facteurs, à savoir la variété des types d'informations présentes dans le dictionnaire, et la latitude laissée dans leur exploitation par les outils d'analyse utilisés par le

péronymiques et méronymiques essentiellement ; la classification ontologique de ces unités lexicales grâce à un étiquetage en classes sémantiques taxinomiques permettant une généralisation simple. On verra plus en détail les particularités de cette ressource dans la section consacrée à son pendant francophone, *EuroWordNet* français (section 3.4.3 page 120).

[19] Il s'agit également d'une ressource lexicale, composée de couches morphologique, syntaxique et partiellement sémantique. Cette dernière est essentiellement composée de classes sémantiques hiérarchiques, ce qui rapproche *AlethDic* de *WordNet*. *AlethDic* est décrit plus précisément à la section 3.4.4 page 125. On trouvera également les raisons qui nous l'avaient fait choisir pour affiner la technique de désambiguïsation sémantique ainsi que celles qui nous ont amené à renoncer à son exploitation dans notre prototype d'enrichissement.

système. Nous avons tenté d'exploiter l'information morpho-syntaxique restrictive fournie par le dictionnaire pour certaines acceptions afin d'imposer ou de filtrer certains sens dans des contextes déterminés. Dans notre cas illustratif, deux circonstances mettent en lumière l'exploitation de ces données morpho-syntaxiques. La première acception principale de l'entrée *coucher* met en jeu la transitivité de *coucher*, ce qui permet d'éliminer les interprétations liées à la construction intransitive de *coucher* (*II.1* à *II.3*) dans les cas où le contexte donne un objet direct à *coucher* (voir 2.12 page 86). De même, la dernière acception principale (*III.1* à *III.4*) se consacre principalement aux acceptions de la construction pronominale de *coucher*. Dès lors, si le contexte d'apparition de *coucher* présente une construction pronominale, c'est au sein de ce sens principal que l'acception correcte se trouve et les autres interprétations peuvent être éliminées. Le principe du filtrage de l'application de certaines règles ne peut toutefois pas être inséré dans les règles elles-mêmes car ce formalisme ne permet pas d'utiliser plusieurs arguments conditionnels, encore moins de manipuler des opérateurs booléens. Les informations supplémentaires que nous exploitons doivent donc appartenir à un ensemble fermé relativement restreint pour créer un ensemble de conditions applicables lors de l'application des règles de désambiguïsation sémantique, et non lors de leur création.

Énoncé :
Elle s'est **couchée** au *chevet* de son compagnon de vie.

Dépendance impliquant le verbe *coucher* :
`SUBJREFLEX(elle,coucher)`
`VMODOBJ(coucher,à+le,chevet)`

Classe sémantique de *chevet* : MEUBLE

Règle de désambiguïsation sémantique appliquée :
```
coucher : VMODOBJ(coucher,à,FORME/MESURE/MEUBLE/CORPS/VEGETAL)
                ==> #coucher.3.1#
```

FIG. 2.16 – Application d'une règle sémantique de désambiguïsation.

L'exemple 2.3.3 illustre le mode de fonctionnement d'une règle de désambiguïsation sémantique : pour une entrée donnée (`coucher`), lorsqu'il y a compatibilité entre une dépendance de l'énoncé et celle qui est présente dans une règle (`VMODOBJ(coucher,à,MEUBLE)`), le numéro de sens correspondant à cette règle (`coucher.3.1`) est sélectionné. Nous avons ici affaire à une **règle sémantique**, qui remplace certaines unités lexicales par leur classe sémantique (`MEUBLE` pour *chevet*). On peut voir combien ce type de règle permet d'élargir les perspectives d'une règle lexicale. En effet, l'énoncé présent ne

présente aucun rapport lexical avec l'exemple utilisé pour créer la règle qui s'est appliquée (*cf.* exemples 2.13 page 87 et 2.15 page 89).

Une dernière perspective de travail concerne des **dépendances syntaxiques** que nous appelons **transitives**. Il s'agit de dépendances dont la définition par l'analyseur syntaxique est multiple, mais qui recouvrent une seule relation syntaxico-sémantique malgré une présentation syntaxique différente (par exemple sujet d'un verbe actif et complément d'agent d'un verbe passif), ou même correspondent à une même relation syntaxique selon la grammaire traditionnelle (par exemple sujet d'un verbe et sujet inverse d'un verbe). Nous avons donc étudié les dépendances syntaxiques extraites par l'analyseur syntaxique pour déterminer celles qui, contenant les mêmes arguments, sont semblables sémantiquement. Les plus représentatives de ces dépendances sont celles qui font intervenir le complément d'agent et le sujet actif, ou l'adjectif attribut et l'épithète, mais de nombreuses autres équivalences sont concernées. En effet, si la présentation phrastique est différente entre chacun des membres de ces couples, le sens des ces groupes n'en reste pas moins le même, et les indices linguistiques permettant une désambiguïsation efficace doivent être relevés. Nous avons donc mis en place une procédure de mise en concordance des différentes relations syntaxiques équivalentes pour générer de nouvelles règles correspondant aux règles de désambiguïsation originelles. Notre cas illustratif nous avait donné une relation de type sujet (voir 2.14 page 88) entre *vent* et *coucher*. Les relations équivalentes sont `INGSUBJ` (sujet du verbe au participe ou au gérondif), `RELSUBJ` (antécédent d'un pronom relatif sujet du verbe), `INVSUBJ` (sujet inversé du verbe) et `PAGENT` (complément d'agent du verbe passif).

Dépendance originelle :

```
SUBJ(vent,coucher)
```

Règle de désambiguïsation issue de cette dépendance :

```
coucher : SUBJ(vent,coucher) ==> #coucher.1.3#
```

Règles générées à partir de dépendances équivalentes :

```
coucher : INGSUBJ(vent,coucher) ==> #coucher.1.3#
coucher : RELSUBJ(vent,coucher) ==> #coucher.1.30#
coucher : INVSUBJ(coucher,vent) ==> #coucher.1.3#
coucher : PAGENT(coucher,par,vent) ==> #coucher.1.3#
```

FIG. 2.17 – Génération de nouvelles règles à partir de relations syntaxiques équivalentes.

La génération des règles s'effectue automatiquement au travers de *XeLDA*. Cette opération est effectuée selon la progression lexicale[20]. Pour chacune des entrées de l'*OHFD*, une consultation du dictionnaire (*lookup*) est effectuée. Chacun des sens est traité séparément et l'information qui lui est propre est analysée pour constituer les règles de désambiguïsation sémantique selon les modalités présentées ci-dessus : analyse syntaxique des exemples et extraction des dépendances ; génération des relations syntaxiques liées aux collocants ; généralisation du lexique en effectuant une consultation sur les classes sémantiques de *AlethDic* ; création éventuelle de filtres liés à l'information morphologique ; génération des relations syntaxiques équivalentes. Tout cela aboutit à la constitution des règles de désambiguïsation sémantique correspondantes.

L'application des règles de désambiguïsation sémantique

Une fois la base de règles de désambiguïsation sémantique constituée, son exploitation est assez simple puisqu'elle est pour ainsi dire symétrique à sa phase de construction. Elle se base en effet sur une analyse du texte à désambiguïser semblable à celle qui a été opérée sur les exemples du dictionnaire. Dès lors, le fonctionnement du module d'application repose sur la plate-forme d'analyse textuelle *XeLDA*, qui sera chargée d'établir les dépendances syntaxiques entre les mots du textes pour les comparer à celles des règles de désambiguïsation. En cas de déclenchement d'une de ces règles, il appartient également à *XeLDA* de fournir l'extrait du dictionnaire correspondant à l'acception sélectionnée. C'est là le fonctionnement de base de cette méthode.

L'exemple 2.18 page ci-contre montre l'application d'une règle de désambiguïsation sémantique. La règle qui s'applique est issue d'une collocation (*cf.* exemple 2.14 page 88). Elle s'applique pour le verbe coucher et vérifie qu'il entretient une relation `PAGENT` avec le mot *pluie* (en tant que relation similaire à une dépendance originale `SUBJ(pluie,coucher)`). Le numéro de sens est le numéro 3.1 (*cf.* figure 2.12 page 86).

Toutefois, cette simplicité apparente masque une certaine complexité dans le déclenchement des règles. En effet, il n'est pas rare – surtout du fait de la généralisation des règles lexicales à travers les classes sémantiques d'*AlethDic* – que plusieurs règles soient susceptibles de convenir à un même contexte. Si ces règles conduisent à une seule et même acception, le problème se résout de lui-même, aucune ambiguïté n'étant maintenue. Dans le cas contraire, un processus de discrimination permet de choisir une seule acception, directement ou par l'attribution d'un poids à chacun des sens

[20] C'est-à-dire que la génération a lieu pour chaque entrée du dictionnaire prise successivement.

Lorsque l'herbe d'alpage n'est plus broutée, elle pousse haut puis elle est **couchée** par la pluie et les intempéries.

Dépendances impliquant *coucher* :
`SUBJPASS(elle,coucher)`
`PAGENT(coucher,par,pluie)`

Règle de désambiguïsation sémantique originale (*cf.* figure 2.14 page 88) :
`coucher : SUBJ(pluie,coucher) ==> #coucher.3.1#`

Règle dérivée appliquée à l'énoncé :
`coucher : PAGENT(coucher,par,pluie) ==>`
`#coucher.3.1#`

FIG. 2.18 – Application d'une règle de désambiguïsation sémantique (règle dérivée d'une collocation).

proposés, selon une méthode que nous allons présenter plus loin.

Tout d'abord, une des qualités qui a suscité le choix de ce dictionnaire réside dans l'ordonnancement statistique des sens de chacune de ses entrées. En effet, l'*OHFD* a été construit à l'aide de très grands corpus[21], et les lexèmes qui y sont décrits voient leurs sens ordonnés selon leur fréquence dans ces corpus. Un aspect de cette méthode de désambiguïsation sémantique est donc basé sur l'hypothèse que de deux sens envisageables dans un contexte et confirmés chacun par une règle de désambiguïsation, c'est celui qui est le plus fréquent qui est probablement le bon. Cette heuristique ne repose bien entendu sur aucun argument linguistique. De plus, aucun indice contextuel n'est susceptible de faire varier l'interprétation d'un mot en fonction de l'environnement dans lequel il se trouve. Elle ne s'affirmera dès lors que dans les cas où aucune autre tentative de désambiguïsation n'aura fonctionné. D'autres méthodes plus contextuelles sont proposées ci-dessous pour résoudre les cas de **désambiguïsations concurrentes**[22].

Un paramètre est particulièrement valorisé lorsque la désambiguïsation standard n'a pas suffi et que le système se trouve dans un cas de désambiguï-

[21] Les corpus qui ont servi à l'élaboration de l'*OHFD* sont des ensembles de textes de plus de dix millions de mots par langue. Ces textes sont généraux et recouvrent un vocabulaire actuel et usuel à l'époque de la rédaction du dictionnaire.

[22] Nous appelons « désambiguïsations concurrentes », ou « désambiguïsations incomplètes » les cas où l'application d'une méthode de désambiguïsation sémantique permet à plus d'un sens de s'affirmer, c'est-à-dire pour cette méthode, lorsque plusieurs règles concurrentes s'appliquent à un même mot dans un même contexte.

sations concurrentes. Il s'agit de la nature de l'information qui est à l'origine de la création de la règle. Une règle lexicale est considérée comme apportant une information sûre, et la distance entre l'information du dictionnaire et le contexte d'application d'une règle qui correspond à ce contexte sera donc nulle (0.0). Une règle lexicale a toujours la prépondérance sur une règle dite sémantique. Toutefois, l'information lexicale est subdivisée en différents types, certains ayant une plus grande importance que d'autres.

La figure 2.19 indique l'ordre d'importance décroissant de chacun de ces types d'information. Le premier type d'information est celui qui a été considéré par les lexicographes comme le plus figé et le dernier comme le moins figé.

règle issue de *collocation*
règle issue d'exemple de type *composé*
règle issue d'exemple de type *idiomatique*
règle issue d'exemple de type *usage*
règle issue d'exemple de type *général*

FIG. 2.19 – Ordre décroissant d'importance des types informationnels lexicaux.

Si le niveau d'importance du type d'information à l'origine des règles qui s'appliquent ne permet pas un choix définitif entre règles lexicales, le déclenchement de plusieurs règles lexicales différentes pour un même sens accorde bien sûr à ce sens un poids proportionnel au nombre de règles qui se déclenchent pour ce sens. Enfin, si aucune de ces deux méthode n'a donné de résultat, le choix s'effectuera dès lors selon l'ordonnancement des acceptions interne au dictionnaire.

L'application des règles dites « sémantiques » implique en revanche une distance d'autant plus importante que le contexte déclencheur est sémantiquement éloigné de l'ensemble de classes propres à la règle. Afin d'effectuer la sélection, la distance entre la liste L1 des classes sémantiques d'une règle potentielle et la liste L2 des classes sémantiques associées à l'argument de la relation en contexte est calculée comme suit :

$$d = \frac{(CARD(UNION(L1,L2)) - CARD(INTER(L1,L2)))}{CARD(UNION(L1,L2))}$$

De la sorte, le choix de la règle se fait lorsque la distance est la moins élevée entre le contexte et l'information du dictionnaire. Cette distance peut varier de 0.0 – lorsque toutes les classes proposées par le contexte sont semblables à toutes les classes proposées par la règle – à 1.0 – lorsque aucune

des classes n'est commune au contexte et au dictionnaire. Dans le cas où plusieurs règles sémantiques désignent le même sens, les distances calculées pour chacune de ces règles sont multipliées entre elles, ce qui a pour effet de réduire la distance globale pour le sens concerné et donc d'augmenter la probabilité de sélection de ce sens. Il peut aussi arriver que des règles sémantiques de même distance avec le texte soient concurrentes. Dans ce cas, c'est la nature de l'information qui préside au choix de la règle selon les modalités présentés plus haut. Si l'information à l'origine des règles concurrentes est de même nature, c'est dès lors l'ordonnancement des sens du dictionnaire qui décide du choix à appliquer. En effet, les sens de chaque mot polysémique dans l'*OHFD* sont ordonnés en fonction de la fréquence de leur représentation dans les corpus qui ont servi à élaborer cette ressource.

Enfin, lorsqu'aucune règle de désambiguïsation sémantique ne peut s'appliquer au contexte du cible, le sens proposé par le système sera le premier sens fourni par le dictionnaire, habituellement l'acception la plus générale, et en tout cas le sens le plus fréquent du mot courant dans les corpus utilisés pour l'élaboration du dictionnaire.

On a déjà pu voir dans l'exposé des différentes méthodes de désambiguïsation sémantique que les démarches lexicales avaient notre préférence dans le cadre de notre méthodologie de traitement de l'information textuelle. La méthode imaginée par Frédérique Segond à XRCE et Luca Dini et Vittorio Di Tomaso au CELI présente non seulement les avantages de ses semblables, mais elle possède également certains atouts qui nous ont amené à la préférer aux autres pour réaliser notre système, bien que nous ayons décidé d'y apporter un certain nombre de modifications dont nous parlerons plus loin (*cf.* section 5.3 page 156).

Les principales qualités des systèmes lexicaux résident dans la qualité de l'information que l'on peut obtenir d'un dictionnaire pour identifier les traits typiques du contexte d'un sens. Un dictionnaire est exploitable comme un corpus étiqueté à la main et vérifié humainement, le plus souvent plusieurs fois. De plus, les dictionnaires présentent habituellement une relative cohérence et homogénéité, ce qui est plus difficilement le cas d'un corpus, par exemple, à moins qu'il n'ait été étudié en profondeur. De tels corpus exploitables sémantiquement sont rares et de dimensions limitées. De plus, il est relativement aisé de changer le dictionnaire d'une méthode lexicale pour en choisir un qui soit mieux adapté aux besoins de la tâche à laquelle on destine le désambiguïsateur sémantique. Enfin, un de nos objectifs, en effectuant une désambiguïsation sémantique étant de relier chaque unité lexicale d'un texte à un maximum d'information la concernant, un dictionnaire est la ressource idéale pour permettre l'identification de cette information.

La méthode de XRCE ajoute à ces avantages l'utilisation soutenue de la syntaxe pour effectuer la discrimination des sens, ce qui permet non seule-

ment d'exploiter une information supplémentaire pour effectuer la tâche de désambiguïsation, mais aussi d'établir des schémas syntaxico-sémantiques lors de l'analyse des textes à désambiguïser, et ainsi préparer un terrain optimal pour vérifier ou infirmer notre hypothèse selon laquelle une telle analyse peut apporter une aide notable pour l'identification et la recherche d'une information dans un texte. Nous notons toutefois que les erreurs qui peuvent survenir au cours de l'analyse syntaxique sont susceptibles de générer du bruit dans les résultats de l'analyse ultérieure effectuée sur les documents. Il importe donc d'utiliser des outils d'analyse les plus performants possible afin d'éviter au maximum ce qui pourrait interférer avec les résultats de la méthode.

2.3.4 L'évaluation dans SENSEVAL et ROMANSEVAL

De même que les conférences MUC et TREC ont répondu à un besoin d'uniformisation des tâches et de l'évaluation des méthodologies proposées en traitement de l'information, SENSEVAL[23] cherche dès sa première campagne en 1999 à définir un protocole d'évaluation des différents systèmes de désambiguïsation, partant de la constatation que les conditions dans lesquelles les différents systèmes présentés dans la littérature étaient jusque là trop différentes pour permettre une comparaison impartiale et juste. SENSEVAL se propose donc de déterminer le cadre définissant la tâche de désambiguïsation sémantique et de fournir un protocole d'évaluation commun sous forme de compétition.

La tâche proposée aux différentes méthodes consiste à associer les occurrences de certains lemmes présents dans un corpus avec leurs sens dans un dictionnaire. Pour la première campagne de SENSEVAL, ce sont les ressources HECTOR qui ont été choisies [Atkins, 1993], composé d'un fragment de 17 Mo du *British National Corpus* (*BNC*) dont l'étiquetage a été réalisé par six lexicographes avec un niveau de convergence de 95 %. Bien que le dictionnaire présente trois niveaux de granularité, c'est le niveau le plus fin qui est retenu pour la tâche de désambiguïsation sémantique.

L'évaluation n'est pas menée sur l'ensemble des mots des textes présentés au système lors de cette première campagne : seules 41 unités lexicales sont concernées (15 substantifs, 13 verbes, 8 adjectifs et 5 mots de catégorie indéterminée)

Lors de la première campagne SENSEVAL, une évaluation analogue a été menée sur deux langues romanes, le français et l'italien, ce qui lui a valu le nom de ROMANSEVAL[24] [Segond, 2000, Calzolari et Corazzari, 2000]. Cette

[23] Les données relatives aux deux campagnes SENSEVAL sont disponibles sur le site http://www.senseval.org.

[24] Voyez http://www.lpl.univ-aix.fr/projects/romanseval/.

évaluation applique les mêmes principes que son homologue anglophone mais, pour le français, exploite le dictionnaire *Larousse* et le corpus du projet ARCADE[25]. Les mots à désambiguïser étaient soixante : vingt substantifs, vingt adjectifs, vingt verbes.

La seconde campagne d'évaluation SENSEVAL menée en 2001 s'est éloignée des principes de la première en s'axant sur une désambiguïsation de l'ensemble des mots du texte, et non plus d'une petite partie prédéfinie d'entre eux. De plus, le dictionnaire dont les sens ont été utilisés comme référence à cette évaluation sont ceux de *WordNet*. Enfin, l'expérience de ROMANSEVAL a été délaissée pour le français, SENSEVAL s'étant étendu à d'autres langues (basque, chinois, tchèque, danois, néerlandais, estonien, italien, japonais, coréen, espagnol, suédois). De ce fait, le système francophone n'a pu participer à cette campagne.

Les critères d'évaluation qui ont présidé aux tests sur le système du français dans [Brun et al., 2001] correspondent à ceux de la campagne SENSEVAL-2, excepté en ce qui concerne l'utilisation des sens de *WordNet*, puisque cette évaluation concernait des textes francophones. Les sens sont ceux du dictionnaire que nous utilisons pour extraire les règles de désambiguïsation sémantique, l'*OHFD* [Corréard et Grundy, 1994]. Les résultats présentés dans cette évaluation sont encourageants (*cf.* tableau 2.1).

Catégorie	Précision Règles lexicales	Précision Règles sémantiques	**Précision toutes règles**	Rappel Règles lexicales	Rappel Règles sémantiques	**Rappel toutes règles**
Noms	0,88	0,48	**0,68**	0,24	0,14	**0,38**
Verbes	0,97	0,50	**0,58**	0,08	0,19	**0,27**
Adj.	1	0,51	**0,68**	0,23	0,23	**0,46**
Total	**0,90**	**0,50**	**0,65**	**0,19**	**0,16**	**0,35**

TAB. 2.1 – Résultats de l'évaluation du système de désambiguïsation sémantique français (extrait de [Brun et al., 2001]).

Suite à cette évaluation, différentes conclusions ont été tirées qui sont remarquables pour nous dans la mesure où cette méthodologie nous intéresse pour le traitement du problème qui nous occupe. Tout d'abord, l'ensemble des erreurs ne sont pas redevables aux règles de désambiguïsation sémantique. En effet, dans certains cas, le défaut provient d'un défaut des analyseurs (morphologique et syntaxique) qui sont appliqués en amont de la désambiguïsation. C'est bien entendu le résultat global qui nous intéresse, mais nous avons vu que nous disposions désormais d'analyseurs plus récents et plus performants.

[25] Voyez http://www.lpl.univ-aix.fr/projects/arcade/.

L'examen des résultats du système de désambiguïsation sémantique pour le français accuse un déficit par rapport à la même méthode appliquée à l'anglais (65 % contre 80 %). Ce déficit est spécialement marqué dans l'utilisation des règles sémantiques (50 % contre 70 %), alors que les règles lexicales restent globalement aussi efficaces (90 %). Le défaut provient donc de la ressource utilisée pour effectuer la généralisation sémantique des règles, *AlethDic*. En effet, de nombreux défauts de cohérence avec le découpage en sens du dictionnaire de référence, l'*OHFD*, ont été constatés, ainsi que l'existence de classes sémantiques trop générales ou couvrant des concepts trop divers.

Enfin, une dernière constatation était liée au problème de la faible couverture. Elle provient essentiellement d'un besoin d'information (exemple ou collocation) pour chaque sens de chaque entrée pour construire une règle de désambiguïsation. Ce besoin n'est pas toujours comblé. Pour information, moins de 40 % des entrées de l'*OHFD* possèdent au minimum une règle de désambiguïsation sémantique. L'information contenue dans le dictionnaire et attachée à un sens particulier de chaque lemme est donc capitale.

Ces remarques tiennent bien compte des réalités auxquelles est confrontée une telle méthode de désambiguïsation sémantique. Il nous faut les prendre en considération lors du choix des ressources lexicales qui y sont adaptées.

2.4 Conclusion

Nous avons besoin d'outils d'analyse textuelle qui permettent d'identifier les mots, les relations entre les mots et le sens de ces mots dans les documents. De plus, les méthodes choisies doivent se prêter à l'étude de textes nombreux et volumineux. Enfin, cette analyse a pour but non seulement la collecte de l'information contenue dans les textes, mais aussi l'enrichissement de cette information. À ce titre, elle se doit d'établir des liens entre les lexèmes et un dictionnaire de référence, garant de la nature et du sens des mots. Par ailleurs, les résultats de l'analyse constituant la structure informationnelle initiale des documents analysés, il faut qu'ils se prêtent à la conservation et à la manipulation de l'information.

Les outils choisis pour effectuer cette analyse correspondent donc bien à ces critères. La chaîne *NTM-XIP* présente la robustesse voulue en même temps qu'une souplesse propre à faciliter l'identification, le stockage et la manipulation de l'information. De plus, la méthode de désambiguïsation sémantique permet d'identifier l'information lexicale liée à chaque sens sélectionné dans tout le dictionnaire choisi comme référence. Il s'agit à présent de déterminer le ou les dictionnaires dont l'information d'une part servira de base à l'analyse linguistique des textes et d'autre part permettra un enrichissement contextualisé et abondant de l'information détectée dans les documents.

Chapitre 3

Les ressources lexico-sémantiques

3.1 Introduction

Pour identifier l'information présente dans une base documentaire et lui donner le plus grand nombre d'apparences possible, il est capital de disposer d'une information lexicale très riche. Cette richesse doit permettre un travail à deux niveaux :

- la richesse syntaxico-sémantique liée au sens des mots favorise le bon fonctionnement de la désambiguïsation sémantique qui doit permettre d'identifier la signification des éléments informatifs qui constituent la base documentaire : c'est le contexte lexico-syntaxique et syntaxico-sémantique qui permet la sélection du sens ;
- la richesse lexico-sémantique permet au système de déduire d'autres présentations d'une information à partir de la forme qu'elle possède dans les textes. Les différentes formes d'une même information sont obtenues à partir d'expressions synonymiques, de dérivés, de méronymes et **holonymes**, d'hyponymes et hypéronymes, de classes sémantiques, etc.

Le choix des dictionnaires est donc particulièrement important car ils déterminent l'exactitude de la structure en fonction de la qualité de leur information, et sa richesse en fonction de la diversité de l'information. Le choix de plusieurs ressources s'imposait pour subvenir à ce besoin de richesse.

Ce chapitre fait une présentation critique de plusieurs ressources lexicales et indique dans quel cadre elles peuvent être utilisées. Nous étudions ainsi le dictionnaire *Dubois et Dubois-Charlier* [Dubois et Dubois-Charlier, 1997], à la couverture lexicale importante et à l'information variée, l'application de génération de dérivés réalisée par Éric Gaussier [Gaussier, 1999], la par-

tie française d'un dictionnaire multilingue à pivot édité par *Memodata*, le *Dictionnaire des synonymes de la langue française* de René Bailly, le réseau sémantique du français de *EuroWordNet* qu'on ne peut décrire sans toucher un mot de son prédécesseur *WordNet*, et la ressource lexicale *AlethDic*.

3.2 Le dictionnaire de *Dubois et Dubois-Charlier*

Notre choix d'une ressource lexicale dans le cadre d'une recherche en structuration de l'information pour l'interrogation d'une base textuelle est guidé par des besoins essentiels. Tout d'abord, le besoin de disposer d'une information aussi exacte que possible, tant au niveau morphologique et syntaxique que sémantique ou même pragmatique. Or ce niveau d'exactitude ne peut être atteint que par une ressource constituée par un lexicographe – ou en tout cas supervisée attentivement par un spécialiste humain. Ensuite, la nécessité de relier des informations de types différents entre elles pour pouvoir déduire d'un indice lié à un seul des types d'information les autres données qui lui sont explicitement liées dans la ressource, et qui peuvent amener à la compréhension d'un mot, d'une expression, d'une phrase.

Le choix de la ressource lexicale est également lié à l'étendue du lexique abordé, au caractère général de ce lexique et à la variété des informations lexicographiques fournies. Ces données linguistiques doivent en effet servir l'approche méthodologique que nous nous sommes fixée. Cette démarche requiert une information morphologique, syntaxique, syntaxico-sémantique et sémantique. Une information pragmatique, ou à tout le moins taxinomique – c'est-à-dire qui classe les concepts et les structure en une hiérarchie sémantique – serait elle aussi la bienvenue et son apport dans certaines situations est assuré. Notons enfin que le choix d'un dictionnaire n'exclut en rien l'exploitation d'une autre ressource dès lors qu'il est possible d'établir une certaine cohérence entre les informations de différentes provenances que nous utilisons.

Le dictionnaire *Dubois et Dubois-Charlier* [1] est une ressource lexicale duelle dans la mesure où elle est composée de deux volumes distincts mais cohérents et complémentaires. La première partie, appelée *Dictionnaire des mots français*, comporte une information sur l'ensemble du lexique, mais cette information est lacunaire en ce qui concerne la catégorie verbale. Les verbes sont en effet traités dans la seconde section, et leur apparition dans le dictionnaire des mots constitue essentiellement un renvoi à ce *Dictionnaire des verbes français*. C'est en effet sur cette catégorie que les auteurs ont choisi de mettre l'accent, et l'information qui est associée aux verbes est par conséquent plus riche que pour aucune autre catégorie grammaticale. Du fait de

[1] Nous le désignerons par la suite plus simplement sous l'appellation « *Dubois* ».

cette plus grande richesse, la structure de l'information verbale est différente de celle qui sert pour les autres catégories, ce qui justifie la dualité de cette ressource. Une autre particularité de ce dictionnaire est de ne traiter qu'une seule acception par entrée. En conséquence, les lemmes polysémiques seront traités sous autant d'entrées qu'ils possèdent de significations, c'est-à-dire que le **traitement** sera **homonymique**.

Il s'agit d'une ressource générale pour le français dotée d'une assise lexicale extrêmement large, puisqu'elle comporte 115 226 lemmes traités sous 140 850 entrées, dont 12 309 lemmes verbaux traités sous 25 609 sens, et donc autant d'entrées. Le *Dubois* comporte plusieurs autres atouts correspondant ou assimilables aux qualités que nous demandons à une ressource lexicale pour effectuer la tâche inhérente à notre recherche. L'examen que nous en faisons ci-dessous le montrera. Du fait de sa dualité informationnelle, nous allons d'abord décrire la partie consacrée à l'information verbale, puis la partie plus générale du dictionnaire. Ensuite viendra notre commentaire sur la ressource.

3.2.1 La partie verbale

L'importance et la diversité des informations contenues dans cette partie verbale justifient l'intérêt que l'on peut porter à cette ressource lexicale. En effet, chacune des entrées comporte pas moins de onze champs informationnels. Ces informations sont lexicales (lemme, niveau lexical), morphologiques (conjugaison, auxiliaire, déverbaux et autres dérivés, adjectifs verbaux, lexèmes dont dérive le verbe), morpho-syntaxiques (construction en expressions composées), syntaxiques (construction syntaxique), syntaxico-sémantiques (construction syntaxique avec sous-catégorisation des groupes, opérateur), sémantiques (domaine, classe, synonymes) et un des champs contient un ou plusieurs exemples. Nous allons examiner chacun de ces champs informationnels.

Lemme Cette rubrique contient principalement l'infinitif du verbe considéré, comme c'est la tradition en lexicographie française. Le lemme est accompagné de la caractéristique syntaxique dominante dans le schème considéré (et présenté dans les autres champs) :

forme simple : *donner 01* ou ***formaliser 02***
forme pronominale : *donner 25(s)*, *donner 27(s'en)* ou ***formaliser 01(s)***
forme négative : *disconvenir (ne)*
forme *être* et participe passé : *enjouer (ê)*

Il peut également être accompagné du complément morpho-syntaxique permettant de replacer le lemme dans une expression idiomatique caractéristique de son utilisation, lorsque seul cet emploi figé du lemme est encore attesté

Lemme	formaliser 01(s)	formaliser 02
Domaine	PSY	MAT
Classe	P1c	T4b
Opérateur	sent offense D	r/d formel
Sens	se choquer,se vexer	donner formalisation à
Exemple	On se f∼ de sa conduite. Cette conduite a f∼ P.	Le mathématicien f∼ une théorie.Cette méthode ne se f∼ pas.
Conjugaison	1aZ	1aZ
Construction	P10b0 T3100	T1308 P3008
Dérivés	1-- -- ----- --	-Q- -- RB-- --
Nom	6L	6L
Niveau	2	5

TAB. 3.1 – Entrée *formaliser* dans le dictionnaire des verbes.

dans ce sens :

expression idiomatique : *férir (sans coup)*

Les lemmes qui comportent plusieurs entrées pour une même forme sont numérotés. Chaque verbe a autant d'entrées qu'il entre dans des « domaines », « classes », « opérateurs » ou « constructions » différentes (*cf. infra*). Dans l'exemple, *formaliser* comporte deux entrées qui divergent au niveau du schème syntaxique (indiqué dans le champ « lemme » et dans le champ « construction »), mais aussi du point de vue des domaine et classe sémantiques, ainsi que pour la dérivation et le registre de langage.

Domaine Ce champ recouvre d'une part (pour les trois premiers caractères) les domaines sémantiques (pragmatiques, techniques ou scientifiques) d'application des entrées verbales (*cf.* la table de ces domaines 3.2 page ci-contre), et de l'autre (le quatrième caractère du code) le registre de langage (*familier*, *populaire*, *littéraire* ou *vieux*) ou les régionalismes (*belgicisme*, *helvétisme* ou *québécisme*). Dans notre exemple, la première entrée de *formaliser* appartient au domaine *psychologique* (PSY) et la seconde au domaine des *mathématiques* (MAT). Les domaines d'application sont, pour les verbes, au nombre de cent vingt-et-un. Il n'a pas été prévu de structurer ces domaines lors de la conception du dictionnaire. Aucune relation n'est donc préétablie entre eux.

Classe Il s'agit de la première rubrique liée directement à la sémantique du verbe. Son codage permet de lui attribuer une classe sémantique générale (premier caractère du code), puis de le diriger vers une sous-classe syntaxico-

ADM [a]	**administration**	**GAT**	**pâtisserie, confiserie**	PAT	pathologie
AER	aéronautique	GEG	géographie	PEA	peausserie, tannerie
ALI	**alimentation**	GEL	géologie	PEC	pêche
ANA	anatomie	GRA	grammaire	PEI	peinture industrielle
ANI	**animal (petit)**	**GRE**	**antiquité grecque**	**PEN**	**justice, peine, délit**
ANT	*antiquité*	HAB	habillement	PET	pétrole
ARB	**arbre, arbuste**	HER	héraldique	PHA	pharmacologie
ARM	arme	HIS	histoire	PHI	philosophie, logique
ART	**arthropode**	HRL	horlogerie	PHN	phonétique
AST	astronomie	HRT	horticulture	**PHS**	**préhistoire**
ATP	**anthropologie**	HYD	hydrologie, liquide	PHT	photographie
AUT	automobile	IND	industrie	PHY	physique
AVI	**oiseau**	INF	informatique	**PIS**	**poisson**
BAC	**bactérie, virus**	INS	instrument, appareil	**PLA**	**plante**
BAT	bâtiment	**ISL**	**islam**	*PLC*	*police*
BIO	biologie	JEU	jeux	POL	politique
BOI	boisson	**JUI**	**judaïsme**	PRE	presse, journalisme
BOT	botanique	LAI	lait	**PSP**	**psychiatrie**
BXA	beaux-arts	LAN	langue	PSY	psychologie
CAN	*chant*	LIN	linguistique	*PTT*	*postes*
CER	céramique	LIT	littérature, textes	QUA	quantitatif
CHI	chirurgie	**LIV**	**livre, reliure**	RAI	rail, chemin de fer
CHM	chimie	LOC	locatif, lieux	REC	récipient, contenant
CHO	**matière**	**LOI**	**loi**	*REL*	*reliure*
CHR	christianisme	**LOQ**	**parole**	**REP**	**reptile**
CHS	chasse	**MAM**	**mammifère**	**RHE**	**rhétorique**
CIR	**cirque**	MAN	manutention	**RLA**	**relation**
CIN	cinéma	MAR	marine	*RLG*	*religion (non chrétienne)*
COL	couleurs, lumière	MAT	mathématiques	**ROM**	**antiquité romaine**
COM	commerce	MEC	mécanique	SER	serrurerie
COS	cosmétologie	MED	médecine	SOC	sociologie
COU	**couture**	MEN	menuiserie, bois	SOM	corps humain, physiologie
CRU	**crustacé**	MES	mesure, métrologie	**SPE**	**spectacle**
CUI	cuisine	MIL	militaire	SPO	sports
CUL	culture, cultivateur	MIN	mines, saline	SYL	sylviculture
CYC	cycles, moto	MOB	mobilier	TAB	tabac
D99	**départements**	**MOL**	**mollusque**	TAU	tauromachie
DAN	danse	MON	monnaie	TEC	technique
DOG	**chien**	**MRL**	**minéralogie**	TEL	télécommunications
DRO	droit, administration	MTA	métallurgie	TEX	textile
ECL	écologie	**MTL**	**métaux**	TIT	titres
ECN	économie	MTO	météorologie	TON	tonnellerie
ECR	écriture, écrits	**MTR**	**métrique**	TOU	tourisme, loisirs
ELT	électricité	MUS	musique	TPS	temps
ELV	élevage	**MYC**	**mycologie**	TRA	travaux publics
ENS	enseignement, pédagogie	**MYT**	**mythologie**	*TUR*	*turf, hippisme*
ENT	**entomologie, insecte**	*NAV*	*navigation, bateau*	TYP	typographie
EQU	équitation, hippologie	NOM	nombre, numération	VAl	vaisselle
ETH	ethnologie	*NUM*	*numismatique*	VEH	véhicule, route
FAC	**façon, manière**	OBJ	objet	VEN	vénerie
FAU	fauconnerie	OCC	occultisme	VER	verrerie
FEO	**féodalité**	OCE	océanologie	**VERB**	**verbe**
FIN	**finance**	**OMB**	**ver, herpétologie**	**VIA**	**viande**
FLO	**fleur**	**OPP**	**opposition**	VIT	viticulture
FOR	fortification	ORF	orfèvrerie	**VOX**	**voix, bruit**
FRI	**froid**	**ORI**	**orientalisme**	ZOO	zoologie
FRO	**fromage**	PAP	papeterie		
FRU	**fruit**	PAR	parenté		

[a] Les codes en **caractères gras** représentent les domaines d'application qui existent dans la partie générale mais pas dans la partie verbale. Ceux en *italiques sans sérif* représentent les domaines d'application qui existent dans la partie verbale mais pas dans la partie générale.

TAB. 3.2 – Table des codes de **Domaine**.

sémantique liée au type (*humain*, *non humain*, *animal*, *non animé*) du sujet ou de l'objet direct et à l'opposition *propre/figuré* (deuxième caractère du code). Dans le cas d'un auxiliaire, la sous-classification est spécifique (`temporalité`, *modalité*, *sujet impersonnel*). Enfin, le troisième caractère prétend spécifier à l'intérieur de chaque classe et sous-classe les constructions syntaxiques typiques de compléments. Dans notre exemple, la première entrée de *formaliser* appartient à la classe `P1c`, qui regroupe les verbes *psychologiques* (`P`) à sujet *humain* (`1`). Le troisième caractère (`c`) code en principe un complément prépositionnel typique régi par la préposition *avec*, mais il semble que cela ne s'applique pas à cet exemple. On verra plus loin que cette information n'est pas fiable. La seconde entrée reçoit le code `T4b`, qui correspond à un verbe de *transformation, changement* (`T`) dont le sujet est de type *non animé* pris dans un sens *figuré* (`4`) et dont le complément prépositionnel, toujours selon la table de résolution des codes, devrait être régi par *de*, ce qui ne s'applique pas non plus à notre exemple. Mais si l'on étudie les différents verbes qui appartiennent à la même classe syntaxico-sémantique que `formaliser 02`, on constate que cette classe est cohérente. En effet, pour chacun de ces verbes, le `b` semble indiquer la présence possible d'un circonstanciel instrumental, c'est-à-dire un groupe prépositionnel introduit par la locution *au moyen* ***de*** ou, plus simplement, par la préposition *avec*. Toutefois, le `b` ne possède pas la même signification dans d'autres classes, mais semble rester cohérent pour chaque classe. La classe syntaxico-sémantique est dès lors pertinente, mais les erreurs d'interprétation de la construction syntaxique (le troisième caractère) ne permettent pas d'envisager son utilisation systématique comme indication de sous-catégorisation sans étudier chaque classe en profondeur.

Opérateur Cette rubrique contient les opérateurs syntaxico-sémantiques et sémantiques qui sous-tendent la définition des classes et l'analyse syntaxique contextuelle du verbe. Il s'agit d'un schéma de l'utilisation du verbe, d'un patron à la fois sémantique et syntaxique. La première entrée de *formaliser* présente un schéma opératoire `sent offense D` qui se traduit par *avoir un sentiment d'offense + complétive impersonnelle introduite par de*. Le schéma de la seconde entrée, `r/d formel`, signifie *rendre ou devenir formel*. On trouvera une liste des opérateurs dans le tableau 3.3 page suivante.

Sens Cette rubrique étroitement liée à la sémantique du verbe contient des synonymes ou **parasynonymes**, ou des définitions abrégées paraphrasant l'entrée. Dans notre exemple, la première entrée présente des parasynonymes (`se choquer, se vexer`), tandis que la seconde donne un court schéma définitoire (`donner formalisation à`) de ce sens de *formaliser*.

abda	enlever à, obtenir de	f.cri	émettre un cri	li	lier qc ou qn à
ag	comportements humains	f.éclat	produire de la lumière	li.accord	mettre en accord qc, qn
aux	auxiliaires et modaux	f.espèce	produire un être vivant	li.clo	fermer
av.car	avoir t. [a] caractère	f.ex	faire sortir de	li.mut	marque un changement
av.som	avoir t. état du corps	f.ire	faire aller qp	li.simul	mettre ensemble
D	impersonnel infinitif en de	f.mvt	faire un mouvement	loq	avoir activité de langage
d	devenir t. (adjectif)	f.op	faire t. opération	loq.bien	dire du bien de
dat	donner qc, qn à	f.sent	donner t. sentiment à qn	loq.mvs	dire du mal de
dat.val	donner force, valeur à qc	f.som	donner t. état du corps	mand	indique une demande
df	défaire une opération	f.son	émettre des sons	mun	doter qn ou qc de
dgrp	défaire ce qui est tenu	f.travail	avoir activité de travail	mut	indique une transformation
dic	communiquer qc, que	ger	diriger qc ou qn	mut.car	indique une modification
dli	défaire ce qui est lié	ger.mens	avoir t. pensée	m.e.accord	mettre en accord juste
dli.accord	défaire un accord	grp	prendre ou tenir	m.e.état	mettre en état qc
dli.clo	ouvrir	grp.mens	avoir t. activité consciente	m.e.marche	mettre en marche qc
dli.simul	défaire ce qui est réuni	grp.mvt	arrêter un mouvement	m.e.mvt	mettre en mouvement
dmu	démunir qn, qc de qc	Inf	impersonnel avec infinitif	m.e. structure	donner une structure à qc
dmut	transformer en sens inverse	ict	donner un coup	percep	avoir t. sensation
d.r/d	inverse de rendre/devenir t.	ind	montrer à	percep.mens	avoir t. connaissance
ê	marque l'existence de	init	commencer une action	Q	impersonnel avec complétive
ex	sortir de	interli	lier des choses entre elles	r	rendre t. (nom)
é.e.état	être dans t. état	ire	aller qp	rag	être ou mettre en morceaux
f	avoir t. activité	lc	être, mettre à t. place	r/d	rendre/devenir t.
fab	réaliser un objet	lc.circ	être, mettre autour	r/d.clair	donner de la lumière à
fac	réaliser une action	lc.per	être, mettre le long	scrut	donner son attention à
fin	achever une action	lc.qp	être, mettre en t. lieu	sent	avoir t. sentiment
f.bruit	émettre un bruit	lc.sr	être, mettre au-dessus de	tact	toucher
f.chant	émettre un chant	lc.ultra	être, mettre au-delà de	val	avoir/prendre la mesure de

[a] On indique *t.* pour *tel, telle*, qui correspond à un adjectif ou un nom dans le champ « opérateur ».

TAB. 3.3 – Liste des « opérateurs ».

Exemple Il s'agit d'une rubrique directement héritée du dictionnaire classique, qui contient une ou plusieurs phrases simples qui illustrent l'emploi de chaque entrée, et notamment le schème syntaxique typique de l'entrée courante. La forme correspondant à l'entrée est symbolisée par son initiale suivie d'un tilde (*f*∼. De plus, la présentation des arguments du verbe répond à des règles particulières. L'objet humain est représenté par le sigle `P` représentant sa catégorie sémantique. Le sujet humain quant à lui est représenté tantôt par le pronom impersonnel *on*, tantôt par un terme générique du domaine pragmatique du verbe ; le verbe dans le sens considéré et son sujet appartiennent donc au même domaine pragmatique. Les sujet ou objet non humains sont eux aussi représentés par un terme générique du domaine pragmatique d'application du verbe. Ainsi, dans la rubrique « exemple » de notre exemple *formaliser*, le sujet humain est représenté tantôt par *on* (première entrée, `On se f~ de sa conduite`), tantôt par le terme générique du domaine des mathématiques *mathématicien* (`Le mathématicien f~ une théorie`), où le « domaine » est `MAT` tant pour *formaliser* que pour *mathématicien*. L'objet humain apparaît sous la forme `P` dans l'exemple de la première entrée de *formaliser* `Cette conduite a f~ P`. Enfin, le sujet et l'objet non humains sont figurés par un terme générique du domaine d'application du verbe : de fait, le terme *conduite*, respectivement sujet (`Cette conduite a f~ P`) et objet (`On se f~ de sa conduite`) dans les exemples de la première entrée de *formaliser*, appartient au domaine *psychologique* (`PSY`) au même titre que cette entrée.

Conjugaison Le champ dédié à la flexion morphologique du verbe est composé de trois caractères. Le premier d'entre eux détermine la catégorie de conjugaison à laquelle appartient le lemme. Selon la répartition établie par [Dubois, 1967], qui se base sur le nombre de radicaux que possède chaque verbe au cours de sa flexion, il existe sept catégories distinctes de verbes. Les deux premières catégories correspondent aux deux premiers groupes traditionnels [Grevisse et Goosse, 1991] tandis que les autres se répartissent les subdivisions du troisième groupe. À l'intérieur de ces catégories, le deuxième caractère détermine quel tableau particulier s'applique au lemme. Le troisième caractère indique lequel des auxiliaires il faut utiliser pour les temps composés.

Construction Cette rubrique définit plus finement et plus précisément que la classe ou l'opérateur la construction syntaxique typique de l'entrée considérée ainsi qu'une certaine sous-catégorisation de ce verbe, c'est-à-dire une définition des arguments qui lui sont propres. Le codage indique d'abord le caractère transitif du verbe (*transitif direct* ou *indirect*, *intransitif*, *pronominal*) puis, en fonction des arguments que la syntaxe du verbe autorise, la nature du sujet et de l'objet (humain, animal ou chose, nombre, complé-

tive), la préposition du complément prépositionnel, la nature du complément circonstanciel (locatif de situation, de direction, d'origine, de transition, temporalité, modalité, causalité, instrumentalité).

La première entrée de notre exemple possède deux constructions typiques : `P10b0` pour une construction pronominale (`P`) à sujet humain (`1`) et sans objet direct (`0`), dont le complément prépositionnel est régi par la préposition *de* (`b`) et pour laquelle aucun complément circonstanciel typique n'est recensé (`0`) ; `T3100` pour une construction transitive directe (`T`) dont le sujet est une chose (`3`) et l'objet un être humain (`1`), mais pour laquelle il n'y a aucun complément prépositionnel (`0`) ou circonstanciel (`0`) typique. La seconde entrée accepte elle aussi deux constructions : `T1308` indique une construction transitive directe (`T`) à sujet humain (`1`), à objet direct chose (`3`), sans indication de complément prépositionnel (`0`) et à complément circonstanciel instrumental (`8`), tandis que `P3008` impose une construction pronominale (`P`) dont le sujet est une chose (`3`), sans objet (`0`) ni prépositionnel (`0`) typiques et dont le circonstanciel est instrumental (`8`).

Dérivation Il s'agit d'une rubrique qui signale l'existence d'adjectifs verbaux et de dérivés nominaux en *-able*, *-é*, *-ant*, *-age*, *-ment*, *-ion*, *-eur*, *-oir(e)*, *-ure* et permet d'en générer la forme exacte à partir du lemme pour les formations régulières, essentiellement grâce à un suffixe. Dans le cas de l'adjectif verbal, le code indique l'existence du positif et/ou du négatif[2]. Les autres codes indiquent la formation par **suffixation**. Dans la première entrée de notre exemple *formaliser*, le code `1` indique un dérivé adjectival en -able dont seul le positif existe : *formalisable*. La seconde entrée possède un dérivé adjectival (`Q`) positif et négatif en *-é* (*[in]formalisé*) coexistant avec le participe passé et l'adjectif de base (*[in]formel*) et une dérivation nominale (`RB`) en *-ateur*/*-ation* : *formalisateur*, *formalisation*.

« Nom » Il s'agit du champ qui permet de retrouver le mot de base (nom ou adjectif) sur lequel le verbe est fondé. Les codes permettent de retrouver ce mot de base malgré la présence de préfixes ou la composition. Le code (`6L`) de notre exemple *formaliser* signifie que l'on peut retrouver l'adjectif de base en *-el* d'un verbe en *-aliser* : *formel* en ôtant 6 lettres au lemme avant de lui ajouter sa désinence adjectivale.

Lexique Ce dernier champ indique le niveau de répertoire lexicographique dans lequel le verbe a été enregistré, ce qui devrait théoriquement indiquer

[2]Par exemple, pour le verbe *adapter*, les codes de dérivation indiquent la coexistence du positif et du négatif pour les adjectifs verbaux en *-able* (*adaptable* et *inadaptable*) et en *-é* (*adapté* et *inadapté*).

une estimation de zone de fréquence ou d'importance de ce mot dans la langue usuelle.

1. Dictionnaire fondamental de 1 500 mots.
2. Dictionnaire de base de 15 000 mots.
3. Dictionnaire contemporain usuel de 35 000 mots.
4. Dictionnaire général de 60 000 mots.
5. Grands dictionnaires de langue ou encyclopédiques, lexiques spécialisés.
6. Recensements personnels.

Cette information nous permet de savoir que le premier sens de *formaliser* est bien plus commun que le deuxième, puisqu'il apparaît dans le dictionnaire de base de 15 000 mots (2) tandis que l'autre n'est signalé que dans des dictionnaires encyclopédiques ou spécialisés (5). Toutefois, la construction de ces dictionnaires n'est pas détaillée, et il est impossible de savoir si ces distinctions fréquentielles sont le fait d'études statistiques sur corpus ou si elles proviennent d'estimations des lexicographes.

3.2.2 Dictionnaire des mots

L'information lexicale de la partie verbale, pour importante qu'elle soit, ne peut à elle seule suffire pour satisfaire les besoins liés à une analyse d'un document. Elle trouve son complément naturel dans la partie générale avec laquelle elle conserve une certaine cohérence, même si les données qui composent cette partie générale sont beaucoup moins riches. Les informations que l'on retrouve dans cette ressource sont surtout morphologiques (pluriel et féminin, dérivations) et sémantiques (domaines d'application) sans pratiquement d'indication syntaxique ou de sous-catégorisation.

La plupart des informations disponibles dans la présente ressource correspondent à celles du dictionnaire des verbes. Notre descriptif de l'information ne s'étendra que sur les différences qui existent entre les deux ouvrages, et n'indiquera que pour mémoire la substance des champs déjà traités. L'information se répartit ici aussi sur onze champs, dont cinq regroupent l'information dérivationnelle.

Lemme Ce champ contient l'unité lexicale selon la forme lemmatisée sous laquelle elle se présente traditionnellement dans les dictionnaires. Les homonymes et les lemmes polysémiques sont numérotés.

lemme numéroté : *social 02*

Les adjectifs sont suivis parfois d'astérisques, une pour indiquer leur antéposition possible lorsqu'ils sont épithètes, deux pour signaler que cette antéposition est obligatoire. Il peut également être accompagné du complément

Lemme	social 01	social 02	sociale (la)
Domaine	SOC	ECN	POLv
Sens	d société	d monde d travail	révolution sociale
Catégorie	A1	A-	-2
Genre - Nombre	ac	ac	-k
Dérivé adjectif	U	-	-
Dérivé nom	B	A	-
Dérivé verbe	B	-	-
Dérivé adverbe	2	-	-
Nom	-	-	I
Lexique	1	1	3

TAB. 3.4 – Exemples d'entrées du dictionnaire des mots.

morpho-syntaxique permettant de replacer le lemme dans une expression idiomatique caractéristique de son utilisation, lorsque seul cet emploi figé du lemme est encore attesté dans ce sens :

emploi figé : *sociale (la)*

expression lexicalisée : *courre (chasse à)*[3]

Il s'agit là des deux seules informations à caractère syntaxique de cette partie du dictionnaire.

Domaine La rubrique « domaine » est sensiblement semblable à celle du dictionnaire des verbes et elle garde une bonne cohérence avec les codes qui la composent (*cf.* table 3.2 page 103). Cependant quelques différences sont à remarquer, comme l'apparition d'un domaine *verbe* (VERB) qui n'a pas de raison d'être dans la section des verbes puisque chaque entrée en est un, et qui permet de distinguer dans la partie générale les entrées traitées dans la partie verbale. Il permet de renvoyer l'utilisateur (humain ou machine) du dictionnaire des mots à une entrée de la partie verbale.

Dans notre exemple 3.4, la première entrée de *social* appartient au domaine *sociologie* (SOC), la seconde au domaine *économie* (ECN) et l'entrée *la sociale* au domaine *politique* (POL).

D'autres domaines font leur apparition soit parce que certains domaines qui n'ont pas d'application verbale peuvent recouvrir des réalités nominales ou adjectivales (le domaine MAM pour *mammifère* ou AVI pour *oiseau*, par exemple), soit pour préciser un domaine qui restait plus général dans la section verbale (notamment les domaines GRE – pour *antiquité grecque* – et

[3]La distinction entre mot composé, collocation et expression idiomatique n'est pas toujours aisée à faire. Le *Dubois* a désigné ces compositions sous le terme « expression idiomatique ». D'aucuns auraient préféré « mot composé ».

`ROM` – pour *antiquité romaine* – dans le dictionnaire des mots précisent `ANT` – pour *antiquité* – du dictionnaire des verbes). Il y a aussi de nouveaux domaines qui ne peuvent être mis en rapport avec les domaines préexistants (par exemple `ADM` pour *administration* ou `FEO` pour *féodalité*).

Ces domaines sont ainsi, pour la section générale, au nombre de 162, mais il faut y ajouter le domaine géographique lié au département. Il se compose d'un `D` auquel s'ajoute le numéro du département concerné, à l'exception du `20` qui désigne les deux départements corses et du `97` qui recouvre tous les départements et territoires d'Outre-Mer. L'information liée au département ne présente que peu intérêt en l'état, car elle n'est pas rattachée au vocabulaire lié au nom du département. Par exemple, il n'est pas possible de savoir qu'un *Grenoblois* est un habitant de l'*Isère*, mais seulement que son département est le *38* (`D38`). Un certain travail est donc nécessaire pour pouvoir exploiter réellement cette information.

Un quatrième caractère peut apparaître dans ce champ, qui donne la possibilité d'indiquer un registre de langage ou un régionalisme selon les mêmes modalités que dans la partie verbale, ou de proposer une majuscule au lemme (`m` impose la majuscule à certaines catégorie grammaticale et `n` impose la majuscule en fonction du sens), ou encore d'indiquer qu'il s'agit d'un nom déposé (`d`). Dans notre exemple 3.4 page précédente, l'entrée *la sociale* appartient à un registre *vieilli* (`v` pour *vieux*).

Sens Comme pour le dictionnaire des verbes, cette rubrique comprend des synonymes ou parasynonymes du lemme dans l'entrée courante (dans notre exemple 3.4 page précédente, le synonyme de *la sociale* est `révolution sociale`), ou une définition abrégée qui paraphrase l'entrée (pour l'entrée *social*, soit `d société` pour « de la société », soit `d monde d travail` pour « du monde du travail »).

Catégorie Cette rubrique n'existe que dans la section générale du dictionnaire car une seule catégorie est présente dans la partie verbale. Elle distingue la catégorie grammaticale des lemmes non verbaux, les verbes étant distingués des autres catégories dans le champ réservé au domaine. Le premier caractère du champ indique la catégorie du lemme s'il n'est ni un nom, ni un verbe :

`A` adjectif	`N` préposition
`K` adjectif sans féminin	`P` préposition et adverbe
`L` adjectif sans masculin	`Q` conjonction
`M` adverbe	`R` interjection

Le second ne concerne que les substantifs. Il permet de valider une certaine sous-catégorisation initiée dans la rubrique « classe » de la section verbale du dictionnaire, à savoir le caractère *animé* ou *non animé*, *humain* ou *animal* des arguments du verbe. C'est avec la présente information que ce codage prend tout son intérêt.

1	*nm* : non animé	6	*nf* : humain féminin
2	*nf* : non animé	7	*n* : animal à deux genres (lion, lionne)
3	*m/f* : non animé à deux genres	8	*nm* : animal masculin
4	*m/f* : humain à deux genres	9	*nf* : animal féminin
5	*nm* : humain masculin		

TAB. 3.5 – Résolution des codes de la sous-catégorisation des noms.

Lorsque le lemme courant appartient à la catégorie verbale, ce champ est laissé vide.

Dans notre exemple 3.4 page 109, la catégorie de la première entrée *social* est soit *adjectif* (A), soit *nom masculin* (nm) *non animé* (1), tandis que la seconde est uniquement *adjectif* (A-). L'entrée *la sociale* est un *nom féminin non animé* (-2).

Genre et nombre Cette rubrique morphologique présente sur un code de deux caractères la formation du féminin et celle du pluriel. Le premier caractère permet de générer la forme féminine du lemme (traditionnellement présenté au masculin singulier) lorsqu'elle existe tandis que le second indique les modalités pour la génération des formes du masculin pluriel. Le féminin pluriel peut être formé par un simple passage au pluriel (*-s*) de la forme féminine générée.

Notre exemple 3.4 page 109 présente une même formation des féminin et pluriel des deux entrées *social* : le féminin se fait simplement par l'adjonction d'un *-e* (a) pour donner *sociale*, et le pluriel est construit par la transformation d'une terminaison *-al*/*-ail* en *-aux* (c). Le féminin pluriel deviendra simplement *sociales*. Quant à l'entrée *la sociale*, son statut de nom féminin ne requiert pas une autre forme féminine (-), mais le code de formation du nombre (k) indique qu'il s'agit là d'un nom possédant un singulier sans pluriel.

Dérivés Les quatre rubriques suivantes concernent la dérivation de mots à partir du lemme. Ces dérivations peuvent être :

1. adjectivales,
2. nominales,
3. verbales,
4. adverbiales.

Comme pour les dérivations basées sur le verbe, les codes de ces champs proposent des suffixes et désinences pour générer à partir du radical du lemme courant les différentes formes qui peuvent en être dérivées.

Dans notre exemple 3.4 page 109, la première entrée de *social* possède un dérivé dans chaque catégorie grammaticale : le dérivé adjectival demande une terminaison en *-iste* (`U`) : *socialiste*, le dérivé nominal une terminaison en *-isme* (`B`) : *socialisme*, le dérivé verbal une terminaison en *-iser* (`B`) : *socialiser* et le dérivé adverbial une terminaison en *-ment* sur la forme féminine du lemme (`2`) : *socialement*. La seconde entrée de *social* ne dispose que d'un dérivé nominal présentant une terminaison en *-ité* (`A`) : *socialité*. L'entrée *la sociale* n'a aucun dérivé.

Lexème dont le lemme est dérivé Ce champ informationnel indique par un code la terminaison du mot dont le lemme courant est dérivé. Le codage indique à la fois la catégorie grammaticale de ce mot et la terminaison permettant de le générer. Seule l'entrée *la sociale* de notre exemple 3.4 page 109 dérive d'un autre mot ; il s'agit d'une base adjectivale se terminant en *-al* (`I`) pour obtenir *social*.

Lexique Comme pour la partie des verbes, il existe dans la section générale un champ indiquant le niveau de répertoire lexicographique dans lequel l'entrée courante est répertoriée. Les lexiques d'origine ne sont cependant pas exactement les mêmes pour les mots et pour les verbes :

1. 4 500 mots
2. 17 000 mots
3. 30 000 mots
4. 50 000 mots
5. 60 000 mots et plus
6. apport personnel

Les deux entrées *social* de notre exemple 3.4 page 109 appartiennent au dictionnaire le plus restreint de 4 500 mots (`1`), tandis que *la sociale* n'apparaît que dans un lexique de 30 000 mots (`3`).

3.2.3 Commentaire

Comme nous l'avons déjà dit précédemment, le *Dubois* est une ressource lexicale de premier plan, tant du fait de sa grande variété d'informations – qui couvrent pratiquement tous les aspects de la langue française écrite – que de l'étendue du lexique traité dans le cadre d'un dictionnaire général. Bien qu'il présente certaines imperfections, ses grandes qualités nous ont amené à opter pour son utilisation comme ressource principale, tout en gardant la possibilité d'exploiter également d'autres ressources lorsque l'information du *Dubois* ne nous paraissait pas à la hauteur de la tâche qui lui était demandée.

Puisque notre démarche repose principalement sur l'identification correcte de la signification des lexèmes et que cette identification s'effectue dans le contexte de la désambiguïsation sémantique lexicale, la répartition des entrées par acception, et non par lemme comme c'est traditionnellement le cas, est un avantage de première importance. En effet, toute l'information du dictionnaire s'en trouve classifiée par sens de mot, et non par mot comme c'est habituellement le cas dans les autres dictionnaires. Cela nous permet de disposer d'indices propres à distinguer le sens correct d'un mot dans un contexte donné au cours de la phase de discrimination de l'acception, et d'autre part de disposer de données linguistiques spécifiques à l'entrée sélectionnée lors de l'enrichissement de la base textuelle.

La cohérence de la ressource est un autre atout important, que même son caractère duel ne peut briser. Notamment, l'information liée au domaine d'utilisation garde son unité malgré certaines différences. Un domaine dans une partie se répartit quelquefois en des sous-domaines plus spécialisés : comme nous l'avons vu, le domaine `ANT` (*antiquité*) dans la partie verbale est subdivisé en `GRE` (*antiquité grecque*) et `ROM` (*antiquité romaine*) dans l'autre section ; il est relativement aisé de gérer cette différence en évitant simplement de descendre au niveau le plus spécifique. De plus, certains domaines inexistants dans une section apparaissent dans l'autre mais il ne s'agit généralement pas d'une entorse à la cohésion, simplement d'un domaine qui n'a pas de raison d'exister dans la section où il n'apparaît pas ; par exemple, les domaines qui marquent l'appartenance à un département (`D08` pour les *Ardennes*) n'auraient aucune justification dans la section verbale.

En outre, le nombre de ces domaines d'application (172 domaines différents) permet une délimitation suffisamment précise du contexte d'utilisation des différentes entrées. Bien que cette information ne fournisse qu'un cadre général au contexte dans lequel le lemme est utilisé – encore ce contexte reste-t-il subjectif[4] – et absolument aucune donnée sémantique directement liée à l'entrée, il s'agit là d'une indication précieuse qui pourra être gérée

[4]Nous avons notamment repéré, sous l'entrée *flinguer 01*, la sélection du domaine *pathologie* (`PAT`). Cette interprétation est discutable.

comme un attribut indicatif par *XIP*, sans forcément lui accorder une valeur contraignante. De plus, l'indication du domaine peut aisément s'insérer dans le dictionnaire utilisé par *NTM* où il deviendra une étiquette.

La rubrique dédiée au « sens », et qui contient des parasynonymes ou des paraphrases de l'entrée est extrêmement précieuse : c'est elle qui nous servira de thesaurus pour faire le lien entre différents mots ou expression de même sens. La paraphrase, qui constitue rarement un mode de définition de l'entrée dans les dictionnaires de langue, constitue en cela un atout original. Cependant, le nombre de ces expressions synonymiques – en moyenne deux par entrée – est généralement trop limité pour suffire à recouvrir tout le champ synonymique d'un mot ou d'une expression, même lorsque cette synonymie dépend du contexte d'apparition dans le cas d'une entrée polysémique. Il nous faudra donc avoir recours à d'autres ressources pour suppléer à cette faiblesse du *Dubois*, tout en conservant l'avantage de la synonymie contextuelle.

Les champs liés à la dérivation, aussi bien les formes dérivées du lemme courant que celle dont est dérivé ce lemme, sont extrêmement rares dans un dictionnaire, où l'étymologie est plus habituellement traitée que le voisinage lexico-morphologique[5]. Or pour une application visant la sémantique lexicale, les informations de dérivation peuvent s'avérer fort précieuses, car elles permettent de construire un lien de sens entre l'acception d'un mot et ses dérivés de catégorie grammaticale différentes, et donc d'une signification seulement approximativement semblable.

Cependant, si la formation de ces dérivés à partir du lemme est souvent aisée à réaliser car bien documentée – indication de la terminaison à retirer pour accoler un suffixe au radical ainsi obtenu – il est aussi de nombreux cas où les instructions ne sont pas suffisamment précises pour générer la forme correcte. Lorsque les consignes lacunaires ne permettent pas la génération des formes correctes, il s'agit habituellement de formations irrégulières ou de transformations du radical. Pour pouvoir utiliser ces informations de manière sure, il faudra donc coupler l'opération de génération avec une vérification lexicale, voire effectuer cette génération à l'aide d'un outil distinct comme celui de [Gaussier, 1999] (*cf.* section 3.3 page 116).

Par ailleurs, et bien qu'un dérivé et le mot dont il dérive aient des acceptions proches dans ce dictionnaire, le remplacement d'une forme de mot par un de ses dérivés d'une autre catégorie grammaticale dans un contexte donné ne peut se réaliser sans bouleversement de la sémantique globale de ce contexte, à moins d'effectuer des adaptations de la structure syntaxico-sémantique de la phrase. Il s'agit dès lors d'étudier les correspondances

[5] Par exemple, il n'existe aucun lien, sémantique ou autre, entre les unités lexicales des différentes catégorie grammaticale dans *WordNet*. Lors du projet *EuroWordNet*, c'est une des remarques faites par [Vossen, 1998].

syntaxico-sémantiques des divers arguments des deux mots concernés. Toutefois, cette étude une fois menée (*cf.* section 4.2.2 page 133), et les problèmes liés à la génération des dérivés résolus, les dérivations devraient permettre d'établir des rapports entre des constructions de nature et de syntaxe différentes mais basés sur des unités lexicale morphologiquement et sémantiquement voisines.

Dans le cadre de la détermination du sens des lexèmes en contexte, deux champs informationnels vont se révéler importants pour notre démarche : il s'agit des informations syntaxico-sémantiques des rubriques fournissant des « exemples » et la « construction » sous-catégorisée. En effet, c'est à partir de ces deux rubriques principalement que nous allons être en mesure de construire des règles permettant de discriminer les sens d'un mot donné, au travers de leurs schémas syntaxiques à toutes deux, grâce aux indications lexicales fournies par les exemples et aux indications de sous-catégorisation des arguments présentés par le champ de construction.

Nous déplorons toutefois de ne pas trouver de données semblables hors de la catégorie verbale, car dès lors, nous ne pourrons générer de règles de désambiguïsation sémantique de ce type pour les autres catégorie grammaticale. De plus, la décision de remplacer dans les exemples le mot-vedette par son initiale suivie du tilde est malheureuse, car elle empêche une simple analyse de la phrase en l'état et impose un pré-traitement de génération avant de pouvoir disposer d'un schéma syntaxique correct, ce qui n'est pas simple à effectuer.

Dans le même ordre d'idée, nous pensions pouvoir utiliser facilement l'opérateur pour obtenir un schéma syntaxico-sémantique strict. Cependant, il est difficile de conserver un niveau de cohérence élevé sur un aussi vaste ensemble, surtout lorsque l'espace réservé à ces données est aussi compact. L'information dont l'opérateur est porteur est importante et très riche, mais sujette à trop de coquilles ou à des abrègements bruts et irréguliers. Dès lors, étant donné le temps qui nous était imparti, nous avons renoncé à l'utiliser, car nous avons constaté qu'il faisait souvent un double emploi avec une information contenue pour partie dans la construction et pour partie dans la classe sémantique.

Nous avons constaté lors de notre étude de cette rubrique « classe » que si les deux premières lettres du code permettent de spécifier une classe sémantique et une sous-classe syntaxico-sémantique, la troisième qui doit sous-catégoriser les compléments prépositionnels ne correspond ni aux attentes, ni au codage défini dans la documentation, qui se contente de renvoyer à la rubrique « construction » sans préciser de quelle manière l'exploiter. Cependant, malgré l'inefficacité de cette dernière information, une sémantique générale du verbe et de ses liens par rapport à ses arguments peut se révéler utile pour comprendre ou paraphraser un texte. Mais une fois de plus, cette

information est réservée aux seuls verbes, ce qui n'empêche pas la classification des autres catégorie grammaticale en contexte dès lors qu'elles sont syntaxiquement liées à un verbe.

Ces possibilités de classification tirent essentiellement leur origine du champ catégorie, qui donne non seulement la catégorie grammaticale de l'entrée dans la partie générale du dictionnaire, mais qui surtout lui confère un certain vernis de sous-catégorisation en indiquant son caractère *humain*, *animal* ou *inanimé*, en parfaite compatibilité avec l'information de la classe verbale. Cette indication permettra en contexte de classifier correctement le verbe et ses arguments syntaxiques.

Les autres champs morphologiques tels que l'indication de formation des genres et nombres ou la conjugaison ne nous seront pas utiles car nous disposons de lexiques permettant non seulement ce type de génération, mais aussi l'analyse des formes fléchies. Ces lexiques, que nous utilisons à travers *NTM* pour l'analyse morphologique, pourraient toutefois être enrichis du vocabulaire supplémentaire présent dans le *Dubois*, mais il s'agit d'un travail relativement long que nous ne pouvions nous permettre d'entreprendre.

Les champs indiquant de quel lexique provient l'entrée semblent indiquer une certaine fréquence de l'acception considérée. Cependant, les dictionnaires utilisés pour définir ces zones fréquentielles sont différents dans les deux sections de la ressource, ce qui est une entorse à la volonté de cohérence constatée jusqu'ici. D'autre part, le codage présent dans cette rubrique est fréquemment soit erroné, soit non documenté, ce qui le rend inexploitable. Enfin, nous estimons que cette information ne repose pas sur une base linguistique suffisante pour pouvoir être exploitée avec un degré de sécurité satisfaisant.

3.3 La morphologie dérivationnelle

Le domaine de la recherche d'information connaît diverses techniques d'enrichissement, centrées essentiellement sur la requête. Notre position qui vise principalement l'enrichissement du texte de la base documentaire interrogée ne nous interdit pas cependant de nous intéresser aux précédentes méthodes avérées. Or la dérivation morphologique des termes de la requête est un type d'enrichissement qui a déjà été testé avec succès.

Cet enrichissement repose sur une constatation : un lexème donné et les dérivés de ce lexème appartiennent ordinairement au même champ sémantique. De la sorte, toute dérivation d'un mot est susceptible d'avoir un sens proche et donc d'être un terme de recherche pertinent dans le cadre d'une requête portant sur la signification dont le mot de base est porteur. Il s'agit dès lors de reconstituer l'ensemble des dérivés de chaque mot.

Notre propos n'est pas de discuter les méthodologies appliquées pour l'apprentissage des règles de dérivation morphologiques et suffixales. Il s'agit généralement de méthodes probabilistes non supervisées dont on peut trouver un aperçu dans [Gaussier, 1999] ou [Snover et al., 2002]. Elles s'appuient sur une description morphologique de la langue observée et étudient de vastes corpus d'où elles extraient un modèle statistique des transformations suffixales constituant un ensemble de règles de dérivation.

Disposant du modèle probabiliste de Éric Gaussier, et de l'outil construit pour générer les dérivés d'un mot donné, nous avons décidé de l'exploiter en en diminuant les contraintes afin d'obtenir le nombre de dérivés le plus grand possible. En effet, le rappel devient notre seule préoccupation dans l'utilisation de cette ressource dès lors que le dictionnaire que nous utilisons nous permet de distinguer les dérivés corrects – ceux qui sont bel et bien issus du mot proposé – des autres, garantissant donc une excellente précision. Cet assouplissement des contraintes modifie du même coup toutes les performances de la ressource employée. Cette possibilité est à la base du choix qui nous l'a fait préférer à d'autres.

Il faut toutefois signaler que nous avons maintenu une exigence dans la génération des dérivés : chacun des termes produits doit obligatoirement apparaître dans le lexique du français pour être validé. Aussi les dérivés abusivement formés du fait de la réduction des contraintes sont-ils filtrés grâce à un dictionnaire général, celui qui permet à *NTM* d'effectuer une analyse morphologique.

Comme nous l'avons vu dans à la section 3.2 page 100, le dictionnaire *Dubois* comporte une information qui nous permet d'éviter de nous soucier de la précision au cours de la tâche de génération des dérivés morphologiques. Un code identifie en effet pour chaque acception d'un lemme le ou les dérivés qui en sont issus. C'est à l'aide de cette information que nous évitons la surgénération après avoir augmenté le rappel en réduisant les contraintes. On peut voir dans l'exemple 3.1 page suivante de quelle manière nous effectuons le filtrage des pseudo-dérivés.

Seules les propositions qui trouvent leur confirmation dans le dictionnaire sont conservées. Les autres sont considérées comme fautives et éliminées. On voit bien par cet exemple l'importance de la suppression des contraintes : si nous les avions conservées, un seul dérivé aurait été trouvé au lieu de six. De plus, dans ce cas précis, le dérivé trouvé est erroné. Une fois les dérivés trouvés, encore faut-il les exploiter à bon escient. Habituellement, on se contente de faire intervenir ces dérivés comme de simples mots-clefs de la requête. Notre approche qui consiste à enrichir le texte plutôt que la requête nous contraint à une méthode plus complexe, mais plus précise. Nous verrons plus loin en quoi elle consiste (*cf.* section 4.2.2 page 133).

Pour le verbe « couper » :

Génération contrainte	Génération libre (non contrainte)	Indication du dictionnaire
coup	coup	–
–	*coupure*	dérivé nominal en *-ure*
–	coupable	–
–	*coupage*	dérivé en *-age*
–	*coupant*	adjectif verbal en *-ant*
–	*coupe*	dérivé nominal (suppr. d'une lettre)
–	*coupeur*	dérivé nominal en *-eur*
–	*coupé*	adjectif verbal en *-é*
–	coupée	–
–	coupon	–
–	couponnage	–

FIG. 3.1 – Filtrage de la morphologie dérivationnelle.

3.4 Les dictionnaires de synonymes

Le dictionnaire *Dubois*, malgré toutes les qualités qui nous ont conduit à sa sélection, comporte toutefois une lacune importante dans le domaine de la recherche d'information. En effet, comme il ne s'agit pas d'un dictionnaire consacré à la synonymie, ce champ informationnel reste relativement restreint. Il ne comporte de fait jamais plus de deux expressions synonymiques par entrée. Nous avons donc jugé que l'utilisation d'un ou plusieurs dictionnaires de synonymes en complément de l'apport du *Dubois* était indispensable, et nous nous sommes intéressé aux dictionnaires de synonymes français disponibles.

3.4.1 Dictionnaire multilingue *Memodata*

Il ne s'agit pas d'un dictionnaire spécialisé dans la synonymie, mais d'une ressource lexicale multilingue traitant cinq langues, dont le français [6]. Comme il s'agit d'un lexique sémantique de traduction basé sur un index qui sert de pivot entre les langues, ce pivot peut toutefois être considéré comme un index de synonymie. En effet, chaque élément de l'index recouvre une signification ce qui permet de naviguer d'une langue à une autre au niveau sémantique. Chacune des langues comprend environ quarante-cinq mille entrées dont chacune comporte sa catégorie grammaticale ainsi que son numéro d'index permettant de faire le lien avec les mots des autres langues de même signification. Ces significations sont au nombre de 37 655.

[6] Les autres sont l'anglais, l'allemand, l'italien, l'espagnol

Étant donné que nous traitons uniquement des documents en français, nous ne nous intéressons bien entendu qu'à la partie française de l'ouvrage et surtout à l'index-pivot qui permet de déceler les unités lexicales porteuses d'une même signification. Même si cette ressource n'est pas d'une grande richesse en ce qui concerne la synonymie, les quelques neuf mille mots partageant un sens avec une autre entrée permettent d'ajouter un certain matériel synonymique à celui dont nous disposons déjà à travers le *Dubois*. Il nous faudra toutefois effectuer un traitement sur cette information, l'identification d'un sens dans le dictionnaire *Memodata* n'étant pas toujours la même que dans le dictionnaire *Dubois*.

3.4.2 Le *Dictionnaire des synonymes de la langue française* de René Bailly

Le *Dictionnaire des synonymes de la langue française*[7] de René Bailly[8] [Bailly et Toro, 1947] dont nous disposons est la version numérisée du dictionnaire papier portant le même nom édité par Larousse. Il se présente sous la forme d'une série d'entrées classées alphabétiquement, mises en correspondances avec leurs synonymes. Ces entrées sont au nombre de 12 738 et l'ensemble des synonymes proposés totalise 28 420 unités lexicales, soit plus de deux synonymes par entrée.

Nous notons après examen de ce dictionnaire de synonymes que dans de nombreux cas il y a implication d'expressions à mots multiples tant dans le champ de synonymie que comme entrée. Cette caractéristique peut se révéler intéressante dans la mesure où nous ne nous intéressons pas seulement à la sémantique des unités lexicales mais également à celle de segments plus importants.

Il faut aussi remarquer que la catégorie grammaticale des entrées n'est pas indiquée dans les champs leur correspondant. Cette particularité peut rendre plus complexe l'exploitation du *Bailly*, car il peut rarement y avoir une équivalence sémantique réelle entre unités lexicales de catégories différentes – excepté bien sûr les cas d'expressions à mots multiples. Il nous reviendra donc de pallier à cette difficulté lors de l'utilisation du dictionnaire, et de restituer sa catégorie grammaticale à chaque entrée. Il faudra aussi, dans le cas des lemmes possédant plus d'une catégorie grammaticale, reconstituer les listes de synonymes propres à chacune des catégories représentées.

[7] Cette ressource nous a été fournie par l'équipe du CRISCO de l'Université de Caen, qui a construit avec Memodata, GREYC et LaTTICe un dictionnaire des synonymes consultable sur http://elsap1.unicaen.fr/dicosyn.html.

[8] Par commodité, nous désignerons à l'avenir sous le nom de « *Bailly* » le *Dictionnaire des synonymes de la langue française* de René Bailly. La version électronique de ce dictionnaire nous a été fournie par Jean-Luc Manguin.

Enfin, nous avons constaté que dans de nombreux cas, les expressions synonymiques sont issues du parler argotique, et que de plus ces unités lexicales sont souvent obsolètes, le parler argotique évoluant parfois très vite. De plus, en parcourant les entrées, nous avons relevé un grand nombre d'erreurs, à moins qu'il ne s'agisse de cas où une signification est tellement désuète que les dictionnaires actuels ne la mentionnent plus. Il y a encore des cas où une entrée appartenant à une catégorie grammaticale reçoit dans ses équivalents synonymiques un ou plusieurs mots appartenant à d'autres catégorie grammaticale, qui sont parfois de simples dérivés du mot-vedette. Ces mots ne peuvent en être les synonymes en aucune façon.

Toutefois, malgré ces défauts nombreux et importants, ce dictionnaire reste essentiel pour notre application étant donné qu'il est le seul dictionnaire spécialement dédié à la synonymie que nous ayons pu nous procurer. D'autre part, nous y avons tout de même trouvé un grand nombre de synonymes intéressants et dont l'exploitation dans le module d'enrichissement pourra se révéler avantageuse dans la mesure où nous comptons appliquer un prétraitement rigoureux à cette ressource du fait de ses nombreuses faiblesses. Ce prétraitement devrait éliminer un grand nombre d'erreurs au prix d'un appauvrissement de la synonymie.

3.4.3 *EuroWordNet* français

La ressource lexicale française de *EuroWordNet* fait partie d'un ensemble de réseaux sémantiques électroniques régissant diverses langues européennes (néerlandais, italien, espagnol, allemand, français, tchèque et estonien) élaboré suivant le modèle anglo-américain *WordNet* mis en œuvre à l'Université de Princeton [9]. Comme nombre de caractéristiques de *EuroWordNet* sont héritées de son prédécesseur *WordNet*, ou découlent d'une critique de celui-ci, il est normal de décrire les principes qui sont à la base de *WordNet* avant de présenter *EuroWordNet*.

WordNet, père d'*EuroWordNet*

WordNet [Miller et al., 1990, Fellbaum, 1998b] s'appuie sur les résultats de recherches psycholinguistiques sur le fonctionnement de la mémoire lexicale humaine [Miller, 1985] : les concepts et unités lexicales sont interconnectés par des relations d'ordre sémantique. Cette constatation a amené ses concepteurs à imaginer une ressource lexicale structurée en terme de sens plutôt qu'en terme de lexèmes. Dès lors, *WordNet* distingue sens de mots, en tant que concept désigné par un lexème, et forme de mot, en tant qu'actualisation physique de la désignation du concept. Une forme de mot peut

[9] Voyez le site http://www.cogsci.princeton.edu/$\sim$wn.

donc avoir plusieurs sens, de même qu'un sens peut avoir plusieurs formes.

WordNet traite séparément les différentes catégorie grammaticale, leur assignant ainsi à chacune un système hiérarchique de classes sémantiques et les structurant par des relations sémantiques. Ces classes et ces structures sont hermétiques, et de ce fait aucune relation ne peut exister entre unités lexicales de classes grammaticales différentes[10]. Quatre réseaux distincts et complètement imperméables sont ainsi disponible : verbes, noms, adjectifs et adverbes. La relation lexicale de synonymie occupe une place prépondérante dans chaque catégorie car elle gouverne toute la structure interne de *WordNet*. Cette relation est en effet constitutive d'ensembles synonymiques appelés *synsets*, et elle est définie relativement à un contexte :

> Deux expressions sont synonymes dans un contexte linguistique C si la substitution de l'une pour l'autre en C ne modifie pas la valeur de vérité de la phrase dans laquelle la substitution est faite ([Miller et al., 1990], page 242).

Les unités lexicales associées par une relation de synonymie constituent un ensemble synonymique. Chaque ensemble synonymique correspond à un sens de mots, et c'est entre ensembles synonymiques que sont établies les autres relations sémantiques exploitées dans *WordNet*. *WordNet* est donc construit comme un réseau sémantique dont les *synsets* sont les nœuds et dont les relations sémantiques sont les arcs. La relation de synonymie s'applique bien sûr à toutes les catégorie grammaticale.

Les relations sémantiques d'hypéronymie et d'hyponymie, aussi appelées relations d'**héritage**, concernent les noms et les verbes. Elles relient un concept général appelé hypéronyme à un concept plus spécialisé, son hyponyme. La relation qui va dans le sens de la généralisation est l'hypéronymie et celle qui va vers la spécialisation est l'hyponymie. L'ensemble de ces relations d'héritage forment la taxinomie implantée dans *WordNet* sous forme d'arbres dont la racine est le terme le plus général, les nœuds autant de concepts plus ou moins spécialisés par rapport à la racine, et les feuilles les entités les plus précises, les plus spécialisées. *WordNet* possède une telle taxinomie uniquement pour les noms et les verbes.

Les relations sémantiques d'holonymie et de méronymie concernent seulement des noms. Elles relient un concept holonyme représentant un tout à un concept méronyme qui constitue une partie du tout. La relation qui va dans le sens de l'entièreté est l'holonymie et celle qui va vers la partie est la méronymie. L'ensemble de ces relations partie-tout forme la **partonomie**, qui est aux relations d'holonymie et de méronymie ce que la taxinomie est aux relations d'héritage. La partonomie est réalisée sous forme d'arbres dont la racine est le terme le plus englobant, les nœuds des sous-parties plus ou moins sub-

[10] À l'exception notable des adverbes, qui sont accessibles par un lien de dérivation à partir des adjectifs.

divisibles et les feuilles les concepts les plus élémentaires, indivisibles. Seuls les noms sont dotés d'arbres partonomiques.

L'antonymie est la seconde relation lexicale présente dans *WordNet*. Elle associe donc deux lexèmes – généralement des adjectifs, mais aussi des noms et des adverbes qui souvent découlent d'adjectifs antonymes ou des verbes [11] – et non deux ensembles synonymiques. Sa définition est complexe, car si elle définit en principe la relation unissant deux unités lexicales décrivant une valeur de vérité inverse (lexème x et lexème *non-x*, *possible-impossible*), elle est couramment appliquée à des unités lexicales qui sont opposées sans pour autant être inverses (*man-woman*, *give-take*).

La **scalarité** est une relation lexicale qui découle de l'antonymie. Deux adjectifs antonymes (*hot-cold*) peuvent en effet appartenir à une échelle de gradation sur laquelle ils se trouvent à des niveaux équivalents, mais opposés (*torrid, **hot**, warm, tepid, cool, **cold**, frigid*). Il s'agit d'adjectifs contraires, et non contradictoires comme ils le sont lorsqu'ils sont antonymes mais n'appartiennent à aucune échelle de gradation.

L'**implication** est une relation sémantique réservée aux seuls verbes. Elle découle de l'implication entre deux propositions contenant les verbes concernés :

> La notion d'implication fait ici référence à la relation qui existe entre deux verbes V_1 et V_2 lorsque la phrase *Quelqu'un* V_1 implique logiquement la phrase *Quelqu'un* V_2 ([Fellbaum, 1998a], p. 77 [12]).

Il faut donc qu'il n'y ait pas de situation concevable dans laquelle la première proposition soit vraie et la seconde fausse. [Fellbaum, 1990] note trois types d'implication verbale : la **cause**, la **présupposition** et la **troponymie**. La cause est une relation d'implication qui relie deux verbes dont le premier est causatif et le second résultatif (*give-have*, donner-avoir). La présupposition relie deux verbes dont l'application du procès du premier implique la réalisation préalable du procès du second (*forget-know*). La troponymie est une relation qui relie deux verbes dont l'un décrit une réalisation particulière du procès de l'autre (*step-walk*, boîter-marcher).

Les relations sémantiques ne sont pas les seules informations assignées aux ensembles synonymiques. Les classes sémantiques sont des étiquettes assignées à chacun des *synset* qui permettent de généraliser les sens de mots recensés dans la ressource et d'établir un rapport de sens entre des ensembles synonymiques qu'aucune des relations sémantiques établies ne permet. Les classes sémantiques sont affectées aux ensembles synonymiques en fonction de leur catégorie grammaticale. Les adjectifs possèdent trois classes séman-

[11] Pour cette catégorie, on l'appelle aussi l'« opposition ».
[12] Par exemple, *He is snoring* (il ronfle) implique *He is sleeping* (il dort).

tiques selon qu'ils sont descriptifs, relationnels [13] ou participiaux. Les classes sémantiques pour chaque ensemble synonymique de noms sont les racines des arbres taxinomiques auxquels ils appartiennent, l'hypéronyme le plus élevé (26 classes). Quant aux verbes, ils possèdent des classes sémantiques différentes, qui se rapprochent de domaines (par exemple CONSUMPTING pour les verbes d'ingestion physique comme *eat* manger et *drink* boire). Elles sont au nombre de 15 : 14 classes pour les actions et événements et 1 pour les états. Enfin, les adverbes ne possèdent pas réellement de classe sémantique. À part les relations de synonymie et d'antonymie, aucune relation sémantique ne les unit, et bien entendu ils n'appartiennent à aucune structure hiérarchique. De ce fait, ils n'appartiennent à aucune classe sémantique dans *WordNet*.

Certaines autres informations apparaissent aussi dans la base de données *WordNet*. On a tout d'abord une courte définition pour chacun des sens de chacune des entrées lexicales accompagnée d'un exemple de l'emploi du lexème dans le sens étudié. On peut également trouver un indice de familiarité de chaque unité lexicale dans une catégorie grammaticale donnée. Cet indice est donné à partir de la polysémie du lexème dans [Hanks, 1986], chaque nouveau sens dans ce dictionnaire incrémentant de un l'indice de familiarité [Tengi, 1998]. Il est toutefois dommage que cet indice de familiarité s'applique à l'unité lexicale plutôt qu'à chaque sens de cette unité.

EuroWordNet, un héritier critique

Le succès remporté par *WordNet* dans le monde anglophone est à l'origine de l'initiative *EuroWordNet* [14] [Vossen, 1998], qui vise à fournir pour plusieurs langues européennes un équivalent de *WordNet*. L'architecture de base de *WordNet* se retrouve ainsi transférée dans chacun des réseaux de *EuroWordNet*, en particulier la notion d'ensemble synonymique ou *synset* dont [Vossen, 1998] donne une définition calquée sur celle de [Miller et al., 1990] (*cf.* page 121) :

> Un *synset* est un ensemble de mots de même catégorie grammaticale que l'on peut interchanger dans un contexte déterminé ([Vossen, 1998], page 73).

Entre ces ensembles synonymiques qui constituent les nœuds du réseau sémantique sont établies des relations d'ordre conceptuel comme on en trouvait déjà dans *WordNet* : hyponymie et hypéronymie, méronymie et holonymie, implication, causalité, etc. La présence de liens hiérarchiques implique l'existence de taxinomies hypéronymiques et méronymiques.

Certaines différences provenant du multilinguisme de ce projet ou de critiques vis-à-vis de *WordNet* existent toutefois dans le projet de conception de

[13] C'est à dire qu'ils découlent morphologiquement et sémantiquement de noms.
[14] Voyez le site http://www.illc.uva.nl/EuroWordNet/.

EuroWordNet. La première est liée à l'**index inter-langue** (ILI), qui permet pour un concept donné de passer d'une langue à une autre. Cet index n'est pas structuré et donne simplement un numéro d'index à chaque ensemble synonymique. La structure reste donc monolingue, ce qui est important pour nous dans la mesure où nous ne traitons que le français.

Les critiques apportées à *WordNet* concernent surtout le manque de domaines, le compartimentage des catégorie grammaticale et la distinction trop fine des sens pour pouvoir être efficacement utilisée notamment en désambiguïsation de sens. Certaines relations sémantique peuvent également être redéfinies pour mieux correspondre à une réalité linguistique ou conceptuelle [Habert, 2000]. On a ainsi certaines relations nouvelles comme *presque synonyme*, *concerné par*, ou une répartition à l'intérieur d'une même relation, par exemple la méronymie qui se subdivise en *localisation*, *fait à partir de*, *membre de* et *partie de*. Notons que le projet *EuroWordNet* ne s'est penché que sur les verbes et les noms.

Cependant, le réseau sémantique français a suivi une voie de construction particulière du fait de son arrivée tardive dans le projet *EuroWordNet*. Du fait du manque de temps, ses concepteurs n'ont pu adapter complètement sa conception aux remarques faites sur *WordNet* et sur les besoins d'une telle ressource. La réalisation du réseau français a donc consisté en une traduction des ensembles synonymiques de *WordNet* 1.5 à l'aide d'une autre ressource, le *Dictionnaire Intégral* DICOLOGIQUE™ [15] [Catherin, 1999]. Le calcul d'une distance entre les *synsets* de *WordNet* et ceux du *Dictionnaire Intégral* déterminent le maintien ou non d'un ensemble synonymique français. Ensuite, une validation manuelle est effectuée, notamment pour la reconstruction de l'hypéronymie, surtout dans le cas de concepts manquants, et de l'antonymie qui doit être intégralement redéployée. L'adjonction de la terminologie informatique est la dernière opération effectuée sur le réseau sémantique français. Il comporte de la sorte 22 745 ensembles synonymiques pour 18 777 entrées lexicales, soit nominales, soit verbales.

Les caractéristiques de *EuroWordNet* français sont très semblables à celles de *WordNet*. Comme pour son prédécesseur anglais, nous déplorons l'absence de domaines taxinomiques dont la présence permettrait une généralisation moins stricte que les relations sémantiques hiérarchiques existantes, ainsi que le manque de données syntaxiques et syntaxico-sémantiques indispensables à une désambiguïsation sémantique correcte. Les informations contenues dans cette ressource sont certes intéressantes, malgré la limitation à deux catégorie grammaticale seulement et l'étroitesse du lexique couvert (moins de 19 000 entrées ne correspondent pas même à un tiers d'un dictionnaire géné-

[15] Le *Dictionnaire Intégral* est une ressource lexicale multilingue de *Memodata* (http://www.memodata.com comportant 120 000 entrées, 25 000 concepts, 350 000 relations, 540 000 mots. Comme *WordNet*, elle privilégie le sens au travers de sa structure en ensembles synonymiques.

ral comme le *Petit Robert* [16]), mais elles ne peuvent intervenir qu'en tant que complément à une autre ressource lexicale qui traite une plus grande part du lexique et fournit une information à la fois morphologique, syntaxique et sémantique.

3.4.4 *AlethDic*, une information importante mais peu cohérente

La société Gsi-Erli [17] a développé naguère une ressource lexicale électronique en langue française dont dispose XRCE. Ce dictionnaire répond aux normes de description d'un *dictionnaire générique* édictées par le projet EUREKA GENELEX [Menon et Modiano, 1993, GENELEX, 1994]. Ce projet a pour but de concevoir le modèle conceptuel d'un fonds lexicographique dont la vocation est de servir de dictionnaire générique pour des applications informatiques traitant le langage naturel. Ces normes prescrivaient un minimum de redondance de l'information, d'où un format lexicologique en couches communicantes depuis le niveau lexical présentant les entrées de la ressource et passant par un niveau morphologique et un niveau syntaxique pour arriver aux données sémantiques, chaque donnée n'étant présente qu'une fois par couche. Le réalisme est également une priorité dans la constitution de cette base lexicale. En effet, l'objectif est de privilégier les besoins des utilisateurs d'un dictionnaire électronique. De ce fait, le projet ne tient aucun compte des théories linguistiques ou lexicologiques plus ou moins émergentes et concurrentes (théorie sens-texte, dictionnaire génératif etc.) et se borne à constituer une base de données lexicale très vaste et très complète. L'objectif de généricité d'un tel dictionnaire le destine à récupérer les informations des divers dictionnaires plus ou moins disparates et plus ou moins spécialisés dans une seule ressource lexicale.

Le dictionnaire *AlethDic* est dès lors une ressource de langue française d'environ 55 000 entrées, ce qui en fait un dictionnaire de taille voisine des dictionnaires généraux classiques. Il est construit à partir des différents lexiques spécialisés que Gsi-Erli a pu ou dû construire au cours des projets déjà menés. Il forme différentes couches selon le niveau linguistique de l'information.

La couche morphologique traite l'information orthographique (variations orthographiques) et phonologique en plus du comportement flexionnel du lemme, et indique ses catégorie et sous-catégorie grammaticales.

La couche syntaxique indique le comportement syntaxique de l'unité lexicale avec laquelle elle est en rapport. Il s'agit d'une sous-catégorisation des groupes syntaxiques reliés à l'entrée : sujet, compléments direct et indirect et

[16] Pour information, nous rappelons que le *Petit Robert* contient 60 000 entrées environ et le *Petit Larousse* 59 000.

[17] Aujourd'hui Lexiquest http://www.spss.com/spssbi/lexiquest/.

prépositionnels pour les verbes, compléments prépositionnels pour les noms, adjectifs et adverbes, complémentation propositionnelle et modes utilisés, rôles de ces éléments composant le schéma syntaxique.

La couche sémantique décrit les unités lexicales grâce à deux artifices. Elle dispose d'abord d'une hiérarchie de classes sémantiques formant une taxinomie et de traits spécifiques qui permettent de raffiner l'information à l'intérieur de la hiérarchie. Ensuite, 192 domaines permettent de relier les substantifs uniquement à l'univers particulier dans lequel ils évoluent.

Malgré toute cette richesse d'information, nous avons décidé de ne pas exploiter cette ressource. En effet, outre la disparité extrême d'*AlethDic* du fait de sa construction à partir d'une multitude de dictionnaires spécialisés, nous rencontrons d'importantes lacunes dans la couche syntaxique du dictionnaire, ainsi que des incohérences dans les schémas syntaxiques de sous-catégorisation. Enfin, c'est de l'information sémantique que nous avons le plus grand besoin. Or le fait de ne pas disposer de la couche sémantique pour les verbes et de ne pas avoir de hiérarchie pour les domaines – exclusivement réservés aux noms, de surcroît – nous amène à rejeter cette ressource par trop hétérogène pour l'utilisation que nous désirerions en faire.

3.5 Conclusion

La richesse et l'architecture homonymique du dictionnaire *Dubois*, ainsi que la variété des informations qui y sont renfermées, nous ont amené à le choisir pour dictionnaire de référence. En effet, les données morphologiques, syntaxiques et sémantiques nous autorisent à l'exploiter tant pour l'analyse des documents que pour leur enrichissement. Toutefois, l'information qui y est rassemblée n'est ni complète, ni exempte d'imperfections.

Nous avons dès lors étudié d'autres ressources lexicales dont l'information est susceptible de compléter les lacunes du *Dubois* ou de corriger ses erreurs. Nous avons sélectionné diverses ressources synonymiques pour enrichir les ensembles de synonymes trop pauvres du *Dubois*, puis un outil de dérivation morphologique capable de pallier les indications parfois déficientes dont nous disposons. Les informations de dictionnaires sémantiques ont également été étudiées, mais nous avons rejeté l'exploitation d'une ressource électronique dont la cohérence et la rigueur étaient trop sujettes à caution.

Ces ressources nécessitent toutefois certains traitements avant de pouvoir être utilisées de manière efficace et cohérente dans les tâches d'analyse ou d'enrichissement auxquelles elles sont destinées. En effet, leur compatibilité avec le dictionnaire de référence doit être assurée pour que les données contenues dans un dictionnaire puissent correspondre à celles des autres ressources.

Chapitre 4

Ajustement des dictionnaires

4.1 Introduction

L'exploitation de plusieurs ressources lexicales dans le cadre d'une même application découle à la fois de la résolution d'un problème de couverture lexicale et de variété de la nature des informations nécessaires présentes dans chaque ressource. Il importe de constituer une seule base de données lexicales contenant l'ensemble des indications voulues que renferme chaque lexique afin de disposer de chaque donnée utile au moment voulu sans parcourir chaque dictionnaire. Cette réalisation se heurte alors au problème de la cohérence entre les différentes ressources.

L'harmonisation entre les ressources est principalement sémantique car le dictionnaire de référence (*Dubois*) effectue un traitement homonymique du lexique, c'est-à-dire que chaque entrée correspond à un sens d'un mot, et non à un lemme. De la sorte, l'information est liée à un sens et non à un lemme comme c'est le cas dans les autres ressources. Il s'agit donc de distribuer l'information contenue dans les lexiques choisis selon le sens des lemmes pour assurer la compatibilité avec le dictionnaire principal. Toutefois, il peut arriver que certaines informations soient contradictoires entre deux ressources. Le choix du *Dubois* comme référence est lié à sa qualité et nous incite à préférer ses choix.

Au cours de ce chapitre, nous nous penchons sur les moyens de remédier aux incohérences entre les dictionnaires qu'il s'agit de corriger. Nous nous intéressons d'abord à la distribution de l'information synonymique par sens, ainsi qu'à la génération, au filtrage et à la répartition des dérivés. Ensuite, nous nous penchons sur la problématique de l'enrichissement d'un lexique morphologique. Enfin, nous évoquons l'opportunité d'intégrer une taxinomie dans ce type d'application.

4.2 Correction de ressources

Les corrections que nous désirons apporter à certaines ressources lexicales ont été abordées dans le chapitre 3 pages 99 à 126 lorsque nous avons signalé au cas par cas les défauts que nous y rencontrions. Il s'agit d'abord de corriger le problème structurel de la distribution des synonymes, liée habituellement au lemme et non aux significations de ce lemme. Ce défaut apparaît dans la liste des synonymes, qui donne une proximité de sens à des mots qui n'en ont pas forcément dans le cas d'une entrée polysémique. L'autre erreur qu'il faut neutraliser a également été signalée auparavant. Nous avons en effet constaté que les informations du *Dubois* permettant de générer les formes dérivées à partir du lemme étaient sporadiquement fausses, ou du moins insuffisamment précises pour permettre la construction directe de ces mots. Nous avons pu remédier à ces problèmes.

4.2.1 Distribution sémantique des synonymes

Au vu de notre méthodologie d'enrichissement de l'information, dont une des caractéristiques prépondérantes consiste à exploiter une synonymie sémantiquement distribuée, le problème le plus saillant que nous ayons à considérer dans les ressources lexicales – excepté le *Dubois* – réside dans le mélange qui est fait des synonymes au sein de chaque entrée sans tenir compte d'une éventuelle polysémie du mot-vedette. Cet enchevêtrement atteint son paroxysme dans le *Bailly*, où aucune distinction de catégorie grammaticale n'est réalisée ni dans les synonymes, ni dans les entrées elles-mêmes. Des ressources comme le dictionnaire *Memodata* ou *EuroWordNet* français lient toutefois les groupes de synonymes à un sens plutôt qu'à un lemme, mais ne distinguent pas toutefois les différents sens d'un même lemme, si ce n'est par un numéro servant de pivot entre les langues.

Or, comme nous l'avons dit, nous nous employons à définir une approche contextuelle de l'enrichissement des textes. Cette approche vise à mettre en correspondance chacune des unités lexicales du texte avec l'ensemble des expressions qui lui sont synonymiques dans le sens que cette unité lexicale présente dans ce contexte. Pour ce faire, à l'intérieur de chaque entrée, il s'agit de définir les ensembles d'expressions synonymiques propres à chaque acception. La distribution des synonymes aura donc lieu non plus en fonction du mot-vedette uniquement, mais bien selon les acceptions de ce mot.

Afin de réaliser cette opération de redistribution synonymique, nous avons décidé d'exploiter l'information de notre ressource de référence, le *Dubois*, car il est le seul dictionnaire dont nous disposons à répartir l'ensemble de son information conformément à la signification des lemmes. Dans le cas présent, l'information que nous pouvons utiliser est sémantique, puisque c'est

une distribution conforme à chacune des acceptions qui doit être effectuée.

Pour ce faire, nous avons mis au point une méthode de filtrage basée sur des informations d'ordre sémantique fournies par notre dictionnaire de référence, le *Dubois*. Ces informations sémantiques sont principalement les domaines d'application (pour l'ensemble du lexique) auxquels s'ajoutent les classe syntaxico-sémantiques (pour les verbes uniquement). Elles permettent de classifier de manière générale [1] les différents sens des entrées polysémiques ou d'étiqueter le sens unique des entrées monosémiques et d'identifier leur sémantique propre.

Entrée	Synonymes	Information sémantique
	Dubois	
ravir01	enlever, retirer	SOCt – S1a
ravir02	dérober, voler	**SOC**t – **S4**a
ravir03	charmer, enchanter	PSY – P2a
	Memodata	
ravir	émerveiller	13520
ravir	enchanter, jeter dans le ravissement, passionner, plonger dans le ravissement	14304
	Bailly	
ravir	charmer, enlever, s'approprier	–
	EuroWordNet	
ravir	exalter, enivrer	@20989@

TAB. 4.1 – Synonymes du lemme « ravir » dans nos différentes ressources.

Face à des ressources lexicales qui proposent des groupes de synonymes liés strictement aux lexèmes sans égard particulier pour les éventuelles variations de sens de ces lexèmes, nous nous trouvons donc à même de constituer des ensembles parmi ces groupes. Ces ensembles sont constitués autour des étiquettes sémantiques dont sont porteuses les unités lexicales synonymiques dans le *Dubois* : le domaine d'application et, dans le cas des verbes, la classe syntaxico-sémantique. Cette particularité permet dès lors d'affecter chaque synonyme d'un mot aux seuls sens de ce mot qui partagent une étiquette sémantique avec le synonyme.

Notre méthode de classification des synonymes s'effectue comme suit. À chacun des candidats synonymes pour un mot donné dans une acception donnée, est associé son domaine dans le *Dubois*, et dans le cas d'un verbe, ses classe et sous-classe sémantiques. Hors du cas particulier des verbes, on

[1] 269 domaines sans relations hiérarchiques, 14 classes syntaxico-sémantiques subdivisées en 54 sous-classes (les 245 sous-sous-classes exclusivement syntaxiques ne seront pas exploitées dans ce cadre).

conserve les candidats synonymes dont le domaine est le même que celui du mot de départ dans le sens considéré. Le domaine d'application est en effet le seul lien sémantique qu'il est possible d'établir entre un mot et son synonyme. Pour la catégorie verbale en revanche, si le domaine d'application apparaît également, la classe sémantique semble dénoter la sémantique de l'entrée elle-même plutôt que son contexte, du moins dans les deux premiers niveaux hiérarchiques, le dernier étant purement syntaxique. Nous avons donc tenté dans un premier temps de sélectionner comme synonymes les candidats présentant de mêmes classes et sous-classes sémantiques que celles du mot de départ dans le sens considéré. Le bilan de ce filtrage était probant en ce qui concernait la qualité de la distribution sémantique des synonymes, mais nous constations un problème de rappel, certains candidats intéressants étant rejetés par ce filtrage. L'examen effectué nous a donc amené à instaurer un autre cas de sélection des candidats synonymes. Nous avons décidé de maintenir également les candidats présentant le même domaine que le mot de départ à condition que, en outre, la classe sémantique (mais pas forcément la sous-classe) soit la même que celle du mot de départ dans le sens considéré. L'examen du dictionnaire ainsi filtré confirme le bien fondé de cette méthode [2].

Le traitement des expressions synonymiques composées de plusieurs lexèmes se démarque toutefois de la méthodologie de filtrage exposée ci-dessus pour contextualiser les synonymes. En effet, nous ne pouvons obtenir une étiquette sémantique cohérente, classe ou domaine selon les cas, pour les expressions à mots multiples car notre dictionnaire de référence, le *Dubois*, traite exclusivement les unités lexicales au travers de ses entrées. Cependant, notre approche vise surtout à restreindre un enrichissement excessif du texte en s'appuyant sur les indices fournis par le contexte. Si la délimitation idéale concerne un enrichissement lié au sens exact d'un lexème dans le texte, il ne s'agit pas de trancher dans les cas où une certaine ambiguïté sémantique se maintient, malgré la désambiguïsation. Cela signifie que plusieurs interprétations sémantiques peuvent être conservées lorsque les indices contextuels ne permettent pas de réduire les hypothèses à une seule acception. Notre stratégie ne nous permet pas de décider du sens qui se rapproche le plus d'une expression synonymique. Aussi avons-nous décidé de conserver les expressions à mots multiples synonymiques d'un mot comme synonymes pour chaque sens de ce mot.

Dans le même ordre d'idée, si le lexème correspondant à un synonyme n'est pas recensé dans le *Dubois* et ne porte de ce fait ni domaine, ni classe sémantique, il sera pareillement versé dans chacun des ensembles synonymiques de l'entrée du dictionnaire de synonymes dans laquelle il apparaît et en portera l'étiquette sémantique. En effet, il n'est pas possible d'affirmer ou

[2] Nous n'avons cependant pas effectué d'évaluation quantifiée de cette approche.

Synonyme	Domaine	Classe	Synonyme	Domaine	Classe	Synonyme	Domaine	Classe
s'approprier	*SOC*	*S*3a	**enlever** (suite)	*SOC*	*S*1a	**retirer** (suite)	COM	E3c
	ECN	U4b		PATt	F1b		DRO	D3f
charmer	PSY	P2a		MIL	S3g		MON	E4b
	OCC	P2c		SOC	S4a		IND	E3c
dérober	MON	D2c		MAN	E3c		ECN	D2e
	*SOC*f	*S*1a		PSYt	P2a		SOC	**S4**h
	SOM	D3f		MUS	R4a		LIT	F4b
	EQU	M1a		LIT	D2c		TYP	R3a
	PAT	M3a		OBJ	D2c		LOC	E3a
	SOC	*S*2b	exalter	LIT	C1i		SOC	E2a
	BAT	R4c		OSY	P1a		LOC	E1a
	LOCp	E1a		PSYt	P2c		MIL	E1a
	PSYp	M2b		SOM	P2c		OCE	E3a
émerveiller	PSY	P2a		COL	M3c	**voler**	MON	D2c
	PSY	P1c		PSY	P1a		MON	N1b
enchanter	OCC	H2h	jeter dans le ravissement	—	—		LIT	**S4**a
	PSY	P1c	passionner	PSY	P1c		SOC	D2c
	PSY	P2c		SOC	P2c		MON	N1b
enivrer	PSY	P1c		PSY	P1a		ZOO	M1a
	BOI	T1b	plonger dans le ravissement	—	—		AER	M3a
enlever	SPO	S3d	**retirer**	JEU	E3c		AER	E3d
	LOC	E3c		LOC	E3c		OBJ	M3a
	TEX	F3c		LOC	E3c		LOC	E1d
	HAB	D2d		HAB	D2d		SOC	E2c
	CHI	D2d		OBJ	D2c		VEH	E3a
	SOC	**S4**a		DRO	S1a			
	ECN	S1a		SOC	**S4**a			
	DRO	S1a		SOC	E2b			

TAB. 4.2 – Étiquetage par le *Dubois* des différents synonymes proposés.

d'infirmer l'appartenance d'un synonyme ou d'une expression synonymique à un ensemble sans disposer du moindre indice concernant son sens.

Une dernière difficulté peut encore apparaître au cours de ce type de traitement de la synonymie. Il est possible qu'un synonyme proposé soit lui-même polysémique, et de ce fait possède plusieurs étiquetages sémantiques distincts. Or dans le cas de l'enrichissement de dictionnaires, la désambiguïsation entre ses différents sens n'est pas possible puisqu'il n'y a pas de contexte permettant d'effectuer un choix parmi eux. Une fois de plus, nous avons décidé de conserver le plus grand nombre de synonymes au détriment peut-être de l'exactitude des ensembles de synonymes. En effet, nous partons du principe que si un des sens du synonyme est considéré comme proche

de celui d'un sens donné de l'entrée, il est probable que ce sens ait provoqué le lien de synonymie. Nous versons donc un synonyme proposé dans un ensemble lorsque un de ses sens au moins présente la même étiquette sémantique que cet ensemble, en estimant que c'est cette signification qui constitue le synonyme de l'unité lexicale de départ.

L'exemple de *ravir* (table 4.1 page 129) illustre bien la difficulté qu'il y a à mettre en concordance toutes les ressources lexicales en ce qui concerne l'information synonymique. À travers cet échantillon, nous montrons facilement en quoi consiste notre méthode de répartition des synonymes proposés. La première démarche consiste, à partir de chacune des propositions de synonyme, à établir un étiquetage sémantique correspondant aux informations du *Dubois* (les domaines et classes sémantiques dans la table 4.2 page précédente). Pour la deuxième acception de *ravir* dans le *Dubois* (table 4.1 page 129, en gras), nous avons un domaine `SOC` pour *sociologie*, et une classe `S4` – nous avons dit que le troisième niveau dans la hiérarchie de classes, exclusivement syntaxique, n'était pas pris en compte – pour *saisir, serrer, posséder* (`S`) avec un **actant** *non animé* dans un sens *figuré* (`4`). Les classe et sous-classe sémantiques permettent de regrouper sous ce sens les propositions de synonymes *enlever, retirer, voler* (table 4.2 page précédente, en gras). Ces synonymes ont été sélectionnés à bon escient, mais d'autres candidats tout aussi valables ne l'ont pas été avec cette première procédure.

L'adjonction d'un processus exploitant les domaines – le même processus que pour les autres catégorie grammaticale – permet d'ajouter aux synonymes *s'approprier, dérober, passionner*. Dans ce cas précis, *passionner* est erroné, et nous notons qu'aucune de ses acceptions ne le relie à la classe sémantique de *ravir* dans le premier sens du *Dubois* (pas de classe en `S`). Nous préférons donc exploiter à la fois le domaine et la principale classe sémantique, ce qui nous amène à conserver parmi les candidats, en plus des synonymes de classe, *s'approprier, dérober* (en caractères obliques), du fait de leur étiquetage semblable que ce soit au niveau du domaine (`SOC`) et de l'appartenance à la même classe (`S`). Les expressions synonymiques à mots multiples sont également conservées. Dans le cas présent, elles sont inexactes, mais ne peuvent être rejetées *a priori*. Elles correspondent au troisième sens de *ravir* dans le *Dubois*.

La figure 4.1 page ci-contre montre les tableaux d'évolution du contenu des dictionnaires lors de la phase de filtrage et également en fonction du type de filtrage appliqué. Il indique aussi la proportion de recouvrement des dictionnaires synonymiques les uns par rapport aux autres.

Potentiel synonymique des ressources avec et sans filtrage :

	Dubois	*EuroWordNet*	*Memodata*	*Bailly*
Total ressource	173 390	26 749	7 450	28 420
Filtrage 1^{er} type	173 390	17 791	5 535	22 157
Filtrage $2^{ème}$ type	173 390	18 811	5 776	22 868

Mesures du recouvrement synonymique au filtrage de 1^{er} type :

	Dubois		*EuroWordNet*		*Memodata*		*Bailly*	
Dubois	173 390	100%	8 231	46,26%	2 944	53,19%	7 109	32,08%
EuroWordNet	8 231	4,75%	17 791	100%	2 106	38,05%	2 971	13,41%
Memodata	2 944	1,70%	2 106	11,84%	5 535	100%	1 102	4,97%
Bailly	7 109	4,10%	2 971	16,70%	1 102	19,91%	22 157	100%

Mesures du recouvrement synonymique au filtrage de $2^{ème}$ type :

	Dubois		*EuroWordNet*		*Memodata*		*Bailly*	
Dubois	173 390	100%	8 665	46,06%	3 102	53,70%	7 399	32,35%
EuroWordNet	8 665	5,00%	18 811	100%	2 204	38,16%	3 124	13,66%
Memodata	3 102	1,79%	2 204	11,72%	5 776	100%	1 149	5,02%
Bailly	7 399	4,27%	3 124	16,61%	1 149	19,89%	22 868	100%

FIG. 4.1 – Filtrage des ressources synonymiques et taux de recouvrement de la synonymie.

4.2.2 Dérivation morphologique pour un enrichissement paraphrastique

Dans notre prospection de techniques permettant de donner à un texte les formes de surface les plus diverses sans en modifier la signification, l'usage de la synonymie est prépondérant. Nous venons de décrire la méthode par laquelle nous entendons améliorer les dictionnaires qui ne sont pas aptes de prime abord à servir notre approche. Un autre procédé proposé pour atteindre notre but consiste à exploiter la parenté sémantique d'une unité lexicale – nous l'appelons « mot original » – avec ses dérivés [Church, 1995, Gaussier et al., 1997, Gaussier et al., 2000]. Cette proximité sémantique a été constatée également dans le *Dubois* qui, pour les lemmes polysémiques, relie les dérivés de la même unité lexicale tantôt à une acception, tantôt à une autre en fonction du sens de chaque forme dérivée.

Cependant, nous avons signalé au cours de l'examen du *Dubois* que l'information destinée à permettre la génération des formes dérivées était occasionnellement erronée ou imprécise. Ces défauts ne remettent pas en cause l'existence de dérivés du type signalé, mais ils empêchent souvent de générer la forme correcte. Pour effectuer cette génération, nous avons donc été

amené à faire appel à l'outil de dérivation morphologique conçu par Éric Gaussier [Gaussier, 1999] dont nous avons décrit les fonctionnalités dans la section 3.3 page 116. Cet outil est capable de nous fournir des formes dérivées avérées, que nous devons redistribuer, voire filtrer, selon les modalités prescrites dans le dictionnaire *Dubois*.

Pour le verbe « couper » :

Formes générées	Instruction *Dubois*	Numéro de sens *Dubois*
coup	–	suppression
coupure	dérivé nominal en *-ure*	1, 7, 9, 10, 12, 14, 16
coupable	–	suppression
coupage	dérivé en *-age*	15
coupant	adjectif verbal en *-ant*	1, 2
coupe	dérivé nominal (– 1 lettre)	1, 3, 9, 19
coupeur	dérivé nominal en *-eur*	1, 29
coupé	adjectif verbal en *-é*	14, 16, 19
coupée	–	suppression
coupon	–	suppression
couponnage	–	suppression

TAB. 4.3 – Génération et distribution ou filtrage par le dictionnaire *Dubois* des dérivés proposés.

L'exemple présenté dans la figure 4.3 illustre bien de quelle manière les indications de la ressource lexicale permettent de filtrer les erreurs de surgénération d'un outil que nous employons sous contrainte minimale, avec pour seule exigence qu'il produise des unités lexicales avérées dans la langue. Ainsi, *coupable* n'est pas retenu par notre filtre car sa génération à partir d'un radical *coup-* et d'un suffixe *-able* ne correspond pas à la réalité de cette unité lexicale. On constate également que les formes dérivées sont distribuées exclusivement sur les acceptions du mot original pour lesquelles le *Dubois* en indiquait l'existence. On peut en effet constater que le mot *coupeur* dérive des sens 1 (synonymes *rompre*, *trancher*) et 29 (synonyme *tailler un vêtement*), mais pas d'autres significations de *couper*, comme par exemple 16 (synonyme *interrompre*) pour lequel *interrupteur* conviendrait mieux. Cette distribution permettra de sélectionner selon leur parenté de sens les dérivés lors de la désambiguïsation sémantique, comme c'est déjà le cas pour les synonymes.

La sélection sémantique des dérivés constitue en soi une amélioration importante de la technique d'enrichissement telle qu'elle est présentée dans [Gaussier, 1999] ou dans [Snover et al., 2002]. Le contrôle de ces dérivés grâce à l'exploitation de l'information d'une ressource lexicale décrivant les relation

du mot original est aussi un perfectionnement notable. Toutefois, ni l'utilisation basique de la dérivation morphologique, ni ces évolutions ne tiennent compte des variations sémantiques qu'un dérivé accuse par rapport à son original. Or un mode d'enrichissement d'un énoncé idéal permet de remplacer dans le texte le segment à enrichir par l'enrichissement qui en découle sans que le sens de l'énoncé n'en soit modifié. Il s'agit donc d'étudier les paramètres susceptibles de modifier la signification de l'énoncé lors de la dérivation et de neutraliser leurs effets. Par exemple, pour un énoncé original *le train entre en gare* et une dérivation *entrée*, la génération d'un énoncé virtuel implique un schéma syntaxique différent pour conserver le sens original : *l'entrée du train en gare*. Cette modification de schéma syntaxique peut provenir soit de l'évolution du sens de l'unité lexicale lors de sa dérivation, soit du changement de catégorie grammaticale lors de cette dérivation.

Identification sémantique des formes dérivées

Dans la section 3.3 page 116 consacrée à la morphologie dérivationnelle, nous avons signalé que le mode de fonctionnement de l'outil de génération des formes dérivées se base exclusivement sur une racinisation (*stemming*) du mot original suivie d'une suffixation. Cette technique de dérivation correspond bien aux indications dérivationnelles du *Dubois*, qui sont suffixales elles aussi. Le *Dubois* fait toutefois une petite entorse à ce principe : il est possible de construire des formes dérivées négatives à partir d'un mot original, grâce à une préfixation. Or l'outil de morphologie dérivationnelle dont nous disposons n'est pas capable d'effectuer cette opération.

Le champ informationnel du dictionnaire *Dubois* prévoit en effet de générer certaines formes négatives à l'aide d'un préfixe *a-* ou *in-*, ou une variation morphologique sur un de ces préfixes. Cette information est cependant insuffisamment précise dans la ressource pour pouvoir être exploitée directement. L'utilisation de l'outil de morphologie dérivationnelle ne permettant pas la préfixation, son exploitation ne pourra remédier au problème dans le cas présent. Toutefois, la sémantique d'une forme négative générée est inversée par rapport à la forme originale, et dès lors seule la négation d'une forme positive permettrait de mettre en rapport la forme positive et la forme négative. Or la grammaire française de *XIP* ne gère pas actuellement la négation. La sémantique des formes négatives est donc difficilement exploitable dans un contexte. Nous reconnaissons toutefois l'importance de cette lacune, qu'il serait intéressant de voir combler.

Le choix à la fois lexical et fonctionnel de baser la dérivation sur la suffixation nous a conduit à étudier les implications sémantiques de cette suffixation. En effet, si les formes dérivées ne présentent pas exactement la même signification que leur mot original, trois paramètres tangibles peuvent

nous guider dans les mécanismes d'évolution du sens : la nature du suffixe utilisé, la catégorie grammaticale du mot original et celle de la forme dérivée.

Dans un premier temps, nous avons cherché à déceler des constantes dans le glissement sémantique qu'implique l'adjonction d'un suffixe à un mot donné. Pour ce faire, nous nous sommes basé sur les observations de [Grevisse et Goosse, 1991] §§168-170 pour l'ensemble des dérivations suffixales proposées par le dictionnaire *Dubois*. Nous avons classifié ces dérivés d'une part selon la catégorie grammaticale du mot original, et de l'autre selon celle de la forme dérivée. Les indications en *caractères obliques* correspondent aux observations où nous nous sommes démarqué de Grevisse.

Formation des dérivés adjectivaux dans la section « verbes » du *Dubois* :

–**able** sert [...] à faire des adjectifs exprimant une possibilité passive (« qui peut être... ») à partir de verbes (*portable*).

–**é** se trouve dans les participes passés, éventuellement employés comme adjectifs (*latinisé*).

–**ant** est la désinence des participes présents, éventuellement employés comme adjectifs (*étincelant*).

Formation des dérivés nominaux dans la section « verbes » :

–**age** pour former des noms indiquant l'action à partir de verbes (*abordage*).

–**ment** pour tirer des verbes [...] des noms exprimant l'action ou le résultat (*abrutissement*).

–**ion** sert surtout à faire des noms d'action à partir de verbes (*abdication*).

–**eur** suffixe ordinaire des noms d'agent (*accompagnateur*). Il sert aussi pour les appareils (*pulvérisateur*).

–**oir** forme des noms désignant des noms d'endroit et des instruments (*abattoir*, *sarcloir*).

–**ure** indique soit une action subie (*contracture*), soit le résultat concret de l'action (*écriture*).

Formation des dérivés verbaux dans la section « mots » :

–**er** a formé et continue de former de nombreux verbes (*abandonner*).

–**iser** a connu un développement considérable en français moderne (*verbaliser*).

–**ifier** s'est surtout développé à l'époque moderne (*densifier*).

Formation des dérivés adjectivaux dans la section « mots » :

–**al** et –**el** pour former des adjectifs dérivés de noms (*frontal*, *industriel*).

–**aire** forme des adjectifs qui ont avec la base des rapports variés (*actionnaire*).

–**ique** pour former des adjectifs, notamment dans la terminologie scientifique et technique (*rabique*).

–**if** forme des adjectifs sur des bases verbales ou nominales (*répulsif*, *narratif*).

–**(i)en** et –**ain** sont devenus des suffixes autonomes marquant l'appartenance (*alsacien*, *diocésain*).

–**able** et –**ible** servent [...] à faire des adjectifs exprimant une possibilité passive (« qui peut être... »). *-ible* est souvent tiré d'un nom en -ion par substitution du suffixe (*perfectible*, *organisable*).

–**ois** et –**ais** se joignent à des noms pour former des noms et adjectifs désignant les habitants ou leur langue (*genevois*, *français*).

–**âtre** a donné des adjectifs exprimant la diminution et l'approximation, souvent avec une nuance péjorative (*blanchâtre*).

–**eux** fournit des adjectifs indiquant une qualité, parfois l'abondance (*courageux*).

–**eur** est le suffixe ordinaire des agents (*pêcheur*).

–**u** forme des adjectifs tirés de noms (*barbu*).

–**esque** sert à former des adjectifs à partir de noms propres, souvent avec une nuance dépréciative (*carnavalesque*).

–**(i)er** forme des adjectifs exprimant une qualité, un rapport (*plaisancier*).

–**in** marque un rapport : ressemblance, matière, origine (*adultérin*).

–**oire** forme des adjectifs tirés de verbes, le plus souvent savant, auxquels correspondent des substantifs en *-tion* (*tentatoire*).

–**ard** forme des adjectifs, souvent avec une nuance péjorative (*fêtard*).

–**uple** *forme des adjectifs et des noms à partir de nombres* (*centuple*).

–**i(a)que** forme des adjectifs dérivés de noms en -ie (*orgiaque*).

–**iste** sert [...] à former des adjectifs indiquant simplement une relation (« qui concerne... ») (*abstentionniste*).

–**ile** *sert à former des adjectifs indiquant la capacité à effectuer une action* (*préhensile*).

Formation des dérivés nominaux dans la section « mots » :

–**(i)té :** les dérivés sont des noms abstraits tirés d'adjectifs (*absurdité*).

–**isme** sert à former des noms masculins, indiquent soit une notion abstraite, soit une doctrine, une activité, une attitude morale ou politique, soit une tournure propre à une langue ou à un parler (*racisme*).

–**ie** suffixe savant, on l'emploie aussi pour des noms de pays et de région (*myopie*, *Wallonie*).

–**at** forme des noms dérivés de verbes pour indiquer une action ou un produit, de noms pour désigner des fonctions (au sens large), parfois le territoire sur lequel elles s'exercent (*électorat*, *actionnariat*).

–**ier** forme des noms désignant des personnes (qui ont une activité en rapport avec la réalité désignée par le mot de base), des contenants, des arbres, des ustensiles divers (*menuisier*, *fraisier*).

–**nce** s'ajoute à des verbes pour former des noms marquant l'action ou son résultat (*abondance*).

–**aie** forme des noms désignant une collection, une plantation de végétaux désignés par la base (*orangeraie*).

–**eur** produit des noms féminins abstraits dérivés d'adjectifs (*blancheur*) ; est le suffixe [...] des noms d'agents et sert aussi pour des appareils (*inspecteur*, *aspirateur*).

–**esse** donne des noms féminins abstraits tirés d'adjectifs (*paresse*).

–**ure** indique soit une action subie, soit le résultat concret de l'action, ou un collectif (*épluchure*, *agriculture*).

–**ard** forme des noms souvent avec une nuance péjorative (*bagnard*, *thésard*).

–**iste** désigne des personnes qui ont une activité, une attitude ou une doctrine en rapport avec la réalité désignée par la base (*pianiste*, *raciste*).

–**ère** forme des noms féminins désignant des personnes (qui ont une activité en rapport avec la réalité désignée par le mot de base), des contenants, des arbres, des ustensiles divers (*étagère*, *ardoisière*).

–**ice** *forme des noms qui désignent le caractère de ce que l'adjectif détermine* (*avarice*).

–**ise** donne des noms abstraits, tirés d'adjectifs (*bêtise*).

–**itude** donne des noms abstraits tirés d'adjectifs ou de noms (*aptitude*).

–**ion** sert [...] à faire des noms d'action à partir de verbes (*réaction*).

–**ment** pour tirer des verbes [...] des noms exprimant l'action ou le résultat (*déménagement*).

–**age** pour former des noms indiquant l'action à partir de verbes (*déballage*).

–**é** *forme des noms en rapport avec le mot d'origine* (*duché*).

–**al** *forme principalement les noms d'alcools à partir d'éléments chimiques* (*chloral*).

–**ade** forme des noms indiquant une action (à partir de verbes) (*bousculade*), un produit, parfois une collection (à partir de noms) (*cotonnade*).

De cette étude, il ressort surtout que les cas sont rares où on peut déduire, même de manière imprécise et sommaire, la signification d'un dérivé à partir de ces seuls paramètres. En effet, nous n'avons pas été capable de définir un panorama cohérent de l'évolution sémantique due à la dérivation suffixale. Il est souvent hasardeux, subjectif ou même contradictoire de donner systématiquement une signification à ces suffixes. L'ajout d'un paramètre supplémentaire dans le cas de dérivations à partir de verbes ou donnant un verbe n'a pas permis de classification plus efficace. De plus, ce sont souvent des considérations historiques, culturelles, sociales apparemment aléatoires, en tout cas difficilement prévisible *a priori* qui interviennent dans l'explication de la formation des dérivés. Tout cela concourt à nous empêcher de déterminer avec suffisamment d'autorité des comportements sémantiques sûrs.

Nous nous sommes dès lors résigné à ébaucher des familles vastes et peu définies à partir des catégorie grammaticale, au sein desquelles la sémantique dérivationnelle reste assez vague et s'appuie essentiellement sur le sens du mot qui en constitue la base.

Nous avons donc abandonné l'idée d'exploiter le suffixe pour définir des schémas d'évolution sémantique de dérivation, car cette information est trop précise et trop diverse à la fois, pour tirer profit de la catégorie grammaticale du mot original et de celle de sa forme dérivée. Ainsi, nous avons :

- origine verbale et dérivé nominal : le dérivé désigne l'action décrite par le verbe ;
- origine verbale et dérivé adjectival : le dérivé a un rapport avec l'action décrite par le verbe ;
- origine nominale et dérivé verbal : le dérivé met en œuvre le concept désigné par le nom ;
- origine nominale et dérivé adjectival : le dérivé a un rapport avec le concept désigné par le nom ;
- origine adjectivale et dérivé verbal : le dérivé met en œuvre le concept que l'adjectif qualifie ;
- origine adjectivale et dérivé nominal : le dérivé désigne le concept que l'adjectif qualifie.
- une origine et un dérivé de même catégorie grammaticale sont considérés comme des formes plus ou moins synonymiques et recouvrant sensiblement les mêmes réalités.

Ces familles de sens très générales nous permettent d'appréhender partiellement la signification des formes dérivées, l'identification du sens des mots originaux donnant lieu à une identification plus précise de la signification de

chacune des unités lexicales générées. On peut dès lors envisager d'utiliser les formes dérivées dans le cadre de l'enrichissement.

Détermination d'un schéma syntaxico-sémantique

La maîtrise de la seule sémantique d'une forme dérivée ne peut toutefois suffire à l'enrichissement d'un énoncé. En effet, pour permettre de mettre en correspondance une question et son élément de réponse présent dans le texte, il s'agit de créer virtuellement un énoncé correspondant au texte original où la forme dérivée prend la place du mot original. Cependant, l'intégration brute de cette forme dans un contexte est susceptible d'altérer fortement le sens premier de ce contexte. L'énoncé virtuel, créé à partir du texte original, devra donc intégrer les différences de surface exigées par la forme dérivée pour maintenir la signification de départ. Ces modifications de surface s'affirment surtout dans les transformations que les relations entre les composantes de l'énoncé original doivent subir pour aboutir à un énoncé virtuel.

Il s'agit donc de modifier le schéma syntaxique et syntaxico-sémantique de la phrase pour que l'intégration de la forme dérivée soit optimale. Pour ce faire, il s'agit d'identifier les schémas syntaxiques typiques que les unités lexicales originelles présentent et de déterminer les transformations que ces schémas subissent lors de la modification de l'énoncé par dérivation de l'unité originale. Dans la perspective du recensement des constantes de modification du contexte syntaxico-sémantique, nous ne pouvons utiliser que les paramètres tangibles qui déjà ont dirigé notre examen de la sémantique de la dérivation suffixale : la catégorie grammaticale du mot original, la nature du suffixe utilisé, la catégorie grammaticale de la forme dérivée. Nous y ajoutons le schéma de sous-catégorisation prescrit par le dictionnaire *Dubois*, déjà utilisé pour les verbes dans la tentative de détermination du sens des dérivés par le suffixe.

Pour effectuer cet examen, nous avons effectué une recherche systématique d'exemples réels sur Internet, considéré pour l'occasion comme un gigantesque corpus. À chacune des combinaisons possibles des paramètres présentés ci-dessus, nous avons pris aléatoirement dans le dictionnaire trois entrées correspondant à ces paramètres (type de dérivation suffixale, catégorie grammaticale originale et dérivée, sous-catégorisation) et nous avons cherché vingt exemples d'utilisation de ces entrées à l'aide d'un moteur de recherche [3]. Nous avons ensuite effectué une analyse syntaxique de ces exemples afin de conserver les dépendances qui concernaient le mot original, puis nous avons remplacé ce mot original par son dérivé et, le cas échéant, modifié la phrase pour qu'elle conserve son sens. Par la suite, nous avons effectué l'analyse syn-

[3] Les vingt premières réponses différentes de Google (http://www.google.fr) ont été retenues pour chaque entrée.

taxique du nouvel énoncé ainsi constitué pour en extraire les dépendances qui concernent le dérivé. Pour les mêmes paramètres, nous avons retenu pour typiques les schémas syntaxiques récurrents présentant cinq occurrences au moins pour chacune des entrées.

Catégorie de l'original : verbe ; catégorie du dérivé : nom.

Suffixe	Sous-cat.	Relation originale	Relation dérivée
-age	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Trans. indirect	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
-eur	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	tout schéma	maintien du schéma
	Trans. indirect	tout schéma	maintien du schéma
-ion	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Trans. indirect	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
-ment	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Trans. indirect	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
		VARG[INDIR](vb,PREP,X)	NMOD[INDIR](nom,PREP,X)
-oir(e)	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Trans. indirect	tout schéma	maintien du schéma
-ure	Transitif	VARG[DIR](vb,X)	NMOD[INDIR](nom,PREP,X)
	Pronominal	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Intransitif	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)
	Trans. indirect	SUBJ(vb,X)	NMOD[INDIR](nom,PREP,X)

TAB. 4.4 – Correspondance des schémas syntaxiques pour les dérivations nominales des verbes.

Nous avons ainsi dégagé de cet examen la méthode de transformation d'une expression en une autre de même sens par le glissement morphologique d'une composante de cette expression. Les relations syntaxiques impliquant les verbes sont principalement celles qui le relient à son sujet (`SUBJ`), à son objet direct (`VARG[DIR]`), à un argument indirect (objet indirect `VARG[INDIR]` ou complément prépositionnel d'un verbe `VMOD[INDIR]`) et à un adverbe (`VMOD[ADV]`). Celles qui mettent en œuvre un nom le relient principalement

Catégorie de l'original : verbe ; catégorie du dérivé : adjectif.

Suffixe	Sous-cat.	Relation originale	Relation dérivée
-able	Transitif	VARG[DIR](vb,X)	NMOD[ADJ](X,adj)
	Pronominal	SUBJ(vb,X)	NMOD[ADJ](X,adj)
	Intransitif	SUBJ(vb,X)	NMOD[ADJ](X,adj)
	Trans. indirect	SUBJ(vb,X)	NMOD[ADJ](X,adj)
-ant	Tous	SUBJ(vb,X)	NMOD[ADJ](X,adj)
-é/-i/	Transitif	VARG[DIR](vb,X)	NMOD[ADJ](X,adj)
-u/-s	Pronominal	SUBJ(vb,X)	NMOD[ADJ](X,adj)
	Intransitif	SUBJ(vb,X)	NMOD[ADJ](X,adj)
	Trans. indirect	SUBJ(vb,X)	NMOD[ADJ](X,adj)

TAB. 4.5 – Correspondance des schémas syntaxiques pour les dérivations adjectivales des verbes.

à un verbe en tant que sujet (`SUBJ`) et objet direct (`VARG[DIR]`) ou indirect (`VARG[INDIR]`). Elles peuvent également le relier à une autre unité syntaxique en tant que tête d'un syntagme prépositionnel (`NMOD[INDIR]`). Enfin, les connexions relatives à l'adjectif sont essentiellement les relations épithètes et attributives (`NMOD[ADJ]` dans les deux cas pour *XIP*).

Les résultats de nos observations sont présentés dans les tableaux 4.4 page précédente et 4.5 pour les dérivations à partir de verbes et 4.6 page ci-contre pour les dérivations à partir des autres catégorie grammaticale. Ces tableaux constituent une concordance entre les schémas syntaxiques identifiés dans les énoncés contenant les mots originaux et les structures syntaxiques correspondantes dans les énoncés modifiés lors du remplacement du mot original par sa forme dérivée. L'information contenue dans la section verbale du *Dubois*, plus riche que dans la partie générale, a permis l'exploitation d'un paramètre supplémentaire (les propriétés syntaxiques des verbes) pour distinguer les différentes possibilités de schémas syntaxiques. Ce paramètre s'est révélé pertinent dans la plupart des cas. Pour les verbes, la nature du suffixe est également un paramètre discriminant entre les différents schémas syntaxiques. Pour les autres catégorie grammaticale en revanche, ni la sous-catégorisation, ni le type de suffixation n'ont permis de distinguer de différences de comportement syntaxico-sémantique au cours de la modification des énoncés.

Les dépendances syntaxiques recensées dans ce tableau correspondent aux relations dégagées de l'examen des énoncés que nous avons définies plus haut. Leurs arguments, qui reprennent en abrégé des catégorie grammaticale en lettres minuscules (`nom`, `adj`, `vb`, `adv`), correspondent au mot original traité ou à sa forme dérivée. Les lettres majuscules `X` et `Y` correspondent à des unités lexicales indéfinies, mais chaque lettre majuscule présente dans

Dérivés verbaux :

Cat. originale	Relation originale	Relation dérivée
Nom	NMOD[INDIR](nom,PREP,X)	VARG[DIR](vb,X)
	SUBJ(X,Y) & VARG[DIR](X,nom)	SUBJ(vb,X)
	SUBJ(X,nom) & VARG[DIR](X,Y)	VARG[DIR](vb,Y)
	SUBJ(X,nom)	SUBJ(X,vb)
Adjectif	NMOD[ADJ](X,adj)	SUBJ(vb,X)
Nom et Adjectif	NMOD[INDIR](nom,PREP,X)	VARG[DIR](vb,X)
	SUBJ(X,Y) & VARG[DIR](X,nom)	SUBJ(vb,X)
	SUBJ(X,nom) & VARG[DIR](X,Y)	VARG[DIR](vb,Y)
	SUBJ(X,nom)	SUBJ(X,vb)
	NMOD[ADJ](X,adj)	SUBJ(vb,X)

Dérivés nominaux :

Cat. originale	Relation originale	Relation dérivée
Nom	tout schéma	synonymie
Adjectif	NMOD[ADJ](X,adj)	NMOD[INDIR](X,PREP,nom)
Nom et adjectif	tout schéma	synonymie
	NMOD[ADJ](X,adj)	NMOD[INDIR](X,PREP,nom)

Dérivés adjectivaux :

Cat. originale	Relation originale	Relation dérivée
Nom	NMOD[INDIR](X,PREP,nom)	NMOD[ADJ](X,adj)
Adjectif	tout schéma	synonymie
Adverbe	VARG[DIR](X,Y) & VMOD[ADV](X,adv)	NMOD[ADJ](Y,adj)
	SUBJ(X,Y) & VMOD[ADV](X,adv)	NMOD[ADJ](X,adj)
Nom et adjectif	NMOD[INDIR](X,PREP,nom)	NMOD[ADJ](X,adj)
	tout schéma	synonymie

TAB. 4.6 – Concordance des schémas syntaxiques pour les dérivations non verbales.

une relation dérivée désigne la même unité lexicale que la même majuscule dans la relation originale. L'argument `PREP` désigne une préposition sur la nature de laquelle nous ne nous prononçons pas. Le sigle `&` représente le *AND* booléen et définit un schéma syntaxique dans lequel deux dépendances sont nécessaires.

Lorsque le mot original et son dérivé possèdent la même catégorie grammaticale, nous n'avons décelé aucun schéma syntaxique qui soit capable de conserver le sens original de l'énoncé par une transformation simple et régulière. Les dérivés de ce type sont en effet presque des synonymes de leur original et seront utilisés au cours de l'enrichissement comme ces synonymes, tout en conservant l'indication de leur origine. Nous avons signalé ces cas par

l'indication `synonymie` dans les tableaux de concordance. Dans certains cas, l'étude des schémas syntaxiques des exemples n'a pas permis de dégager de constante dans la modification de la structure. Cependant, pour éviter de perdre le bénéfice d'un enrichissement possible, nous avons choisi de conserver la forme dérivée sans lui adjoindre de schéma d'évolution syntaxique. Nous conservons toutefois l'information que cette forme ne convient pas au contexte syntaxique de l'énoncé dans lequel elle peut être placée. Nous avons indiqué cette lacune syntaxique par la mention *maintien du schéma*.

Toutes ces données peuvent être intégrées dans les ressources lexicales à l'intérieur des entrées et selon les sens définis au départ dans le champ « dérivation » du dictionnaire *Dubois*, afin d'être directement exploitables lors des phases d'analyse et d'enrichissement. Nous verrons dans le prochain chapitre, consacré à la réalisation d'un enrichissement de texte, de quelle manière les corrections apportées au dictionnaires peuvent être intégrées à notre démarche.

4.3 Élargissement informationnel des ressources lexicales

Chacun des dictionnaires utilisés par notre système a maintenant reçu les corrections qui lui étaient nécessaires. Nous devons à présent nous intéresser au processus d'élargissement de l'information de certaines de ces ressources au moyen de données présentes dans d'autres dictionnaires. En effet, pour éviter d'avoir recours plusieurs fois à un même dictionnaire au cours de l'analyse séquentielle, chaque ressource sera disponible une seule fois mais l'ensemble de l'information lexicale utile sera disponible dans chaque dictionnaire.

L'élargissement informationnel de ressources que nous devons réaliser concerne essentiellement deux des dictionnaires : le lexique utilisé par *NTM* pour effectuer le découpage en mots et l'analyse morphologique, auquel il nous faut apporter l'information sémantique provenant des champs « classe » et « domaine » du *Dubois*, et le dictionnaire *Dubois* lui-même, qui ne dispose pas d'une hiérarchie sémantique taxinomique. Le premier de ces compléments est requis par le système de désambiguïsation sémantique pour permettre le fonctionnement des règles sémantiques. Le second doit permettre une généralisation des unités lexicales plus ou moins importante qui peut servir pour la mise en correspondance des questions et réponses lors de la phase d'interrogation.

4.3.1 Ajout de sémantique dans le lexique morphologique

Lorsque nous avons décrit la méthode de désambiguïsation sémantique (*cf.* section 2.3 page 75) que nous avons choisie pour identifier la signification des éléments qui composent les textes à interroger, nous avons notamment décrit le fonctionnement des règles permettant la discrimination des sens des unités polysémiques. Or si certaines de ces règles étaient *lexicales*, et donc fonctionnaient grâce à l'identification des mots qui constituent le contexte de l'unité à désambiguïser, il en est d'autres que nous appelions *sémantiques*, axées sur l'appartenance des lexèmes du contexte à des groupes sémantiques, les classes d'*AlethDic* pour le français ou celles de *WordNet* pour l'anglais.

Pour permettre à ce type de règles de fonctionner aussi dans notre système, qui n'exploite pas la ressource *AlethDic*, il est important de fournir au désambiguïsateur sémantique l'information dont il a besoin sur la nature sémantique des mots à désambiguïser. Or l'étape d'analyse morphologique est la seule à effectuer une recherche dans un lexique sur tout le dictionnaire avant l'étape de désambiguïsation sémantique. Il est donc logique d'exploiter cette phase lexicale pour distribuer l'information sémantique nécessaire.

L'information sémantique que nous devons apporter au lexique morphologique provient du dictionnaire *Dubois*. Il s'agit des domaines d'application (pour l'ensemble du lexique) et des classes sémantiques (pour les verbes uniquement). Le lexique morphologique ne distingue pas les différents sens des lexèmes. Il n'en a pas besoin, aucune différenciation sémantique ne s'effectuant à ce niveau d'analyse. Chacune de ses entrées reçoit donc l'ensemble des étiquettes sémantiques correspondant au lemme de cette entrée. La distinction des sens et donc l'élimination des étiquettes erronées interviendra lors de la phase de désambiguïsation sémantique.

Pour illustrer l'adjonction de sémantique dans le lexique morphologique, nous avons présenté (*cf.* figure 4.7 page suivante) la forme *commence* de l'exemple d'analyse de *NTM* 2.2 page 59. Le lexique morphologique présentait cinq possibilités d'interprétations morphologiques de cette forme de mot. Or le dictionnaire *Dubois* comporte sept entrées de *commencer* qui présentent une combinaison différente domaine-classe. Chacune des analyses de *commence* est donc multipliée par sept, une par combinaison domaine-classe. Le lexique morphologique comporte donc maintenant 35 entrées, dans lesquelles les domaines sont signalés par un préfixe `DOM_` et les classes par un préfixe `CLA_`. À travers l'analyse morphologique de la forme *commence* par *NTM* (figure 4.7 page suivante), on peut voir que l'information du lexique morphologique a été élargie et que les étiquettes morphologiques sont affectées à chacune des propositions d'analyse. Ces étiquettes sémantiques sont assignées à la forme de mot durant l'analyse morphologique jusqu'à la décision du système de désambiguïsation sémantique.

fonctionnaire	fonctionnaire	+InvGen+SG+DOM_ADM+human+Noun
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_TPS+CLA_X4a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_TPS+CLA_M4b+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_TPS+CLA_X1a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_ENS+CLA_M2c+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_VEH+CLA_L3a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_TEC+CLA_R3a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P1+DOM_PAT+CLA_M4b+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_TPS+CLA_X4a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_TPS+CLA_M4b+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_TPS+CLA_X1a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_ENS+CLA_M2c+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_VEH+CLA_L3a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_TEC+CLA_R3a+Verb
commence	commencer	+avoir+parSN+IndP+SG+P3+DOM_PAT+CLA_M4b+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_TPS+CLA_X4a+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_TPS+CLA_M4b+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_TPS+CLA_X1a+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_ENS+CLA_M2c+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_VEH+CLA_L3a+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_TEC+CLA_R3a+Verb
commence	commencer	+avoir+parSN+Imp+SG+P2+DOM_PAT+CLA_M4b+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_TPS+CLA_X4a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_TPS+CLA_M4b+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_TPS+CLA_X1a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_ENS+CLA_M2c+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_VEH+CLA_L3a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_TEC+CLA_R3a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P1+DOM_PAT+CLA_M4b+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_TPS+CLA_X4a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_TPS+CLA_M4b+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_TPS+CLA_X1a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_ENS+CLA_M2c+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_VEH+CLA_L3a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_TEC+CLA_R3a+Verb
commence	commencer	+avoir+parSN+SubjP+SG+P3+DOM_PAT+CLA_M4b+Verb

TAB. 4.7 – Exemple d'analyse de *commence* par le lexique morphologique après son élargissement sémantique.

Il faut cependant remarquer que lorsque le vocabulaire du lexique morphologique et celui du dictionnaire *Dubois* ne coïncident pas, aucun élargissement ne peut avoir lieu si le dictionnaire est plus étendu, mais aucun retrait ne peut se produire si le dictionnaire est lacunaire. De fait, lorsqu'un lemme du lexique est inconnu du *Dubois* ou que ses caractéristiques morphologiques sont différentes, aucun étiquetage sémantique n'est ajouté dans le lexique.

À l'inverse, si un mot du *Dubois* est absent dans le lexique, cette entrée du *Dubois* n'est pas ajoutée au lexique malgré l'information morphologique présente dans le dictionnaire. En effet, cette opération d'insertion de nouvelles unités morpho-lexicales demande un important travail de reconstruction du transducteur qui constitue le lexique morphologique. Dans le cadre de cette recherche, il est irréaliste de s'engager dans une entreprise de cette ampleur, malgré l'intérêt que cela présente.

L'information sémantique liée aux domaines d'application et aux classes sémantiques est exploitée dans le cadre d'un type de règles de désambiguïsation déjà connu et appliqué dans le système développé à XRCE (Xerox Research Centre Europe). Toutefois, au cours de notre examen des ressources lexicales dont nous disposons, nous avons signalé une autre information d'ordre syntaxico-sémantique qui pourrait donner lieu au développement d'un nouveau genre de contrainte contextuelle et donc d'un nouveau type de règles de désambiguïsation sémantique. Il s'agit des schémas syntaxiques de sous-catégorisation.

Sans présumer de l'intérêt de cette information dans la cadre de la discrimination de l'acception correcte d'un mot polysémique en contexte (le travail de sélection du sens des mots et son importance dans notre démarche seront abordés plus loin dans la section 5.3 page 156), le simple fait qu'elle ait retenu notre attention lors de l'examen du dictionnaire la rend susceptible d'être exploitée ultérieurement pour le traitement que nous avons à effectuer sur les documents. Nous avons dès lors décidé d'anticiper sur les besoins éventuels du traitement de la sémantique des lexèmes et d'intégrer l'information sémantique correspondant aux schémas de sous-catégorisation au lexique d'analyse morphologique.

Cette information se présente sous la forme de trois catégories sémantiques, limitées aux seuls substantifs : *humain*, *animal*, *inanimé*. Leur implantation dans le lexique d'analyse morphologique est semblable à celle du précédent étiquetage sémantique et est effectuée en même temps que cet élargissement de l'information dans le transducteur. En effet, lors de chaque addition d'une information liée à un nom présent dans le *Dubois*, il suffit d'ajouter non pas une étiquette correspondant au domaine d'application du lexème dans le sens visé, mais deux étiquettes, la première correspondant à ce domaine, la seconde au trait sémantique. Dans le cadre des noms monosémiques, ces information sont donc simplement ajoutées, tandis que les entrées polysémiques multiplient, comme précédemment, chaque proposition d'analyse par le nombre d'acceptions existant pour cette entrée.

L'exemple 4.7 page ci-contre montre deux cas d'analyse morphologique, l'une d'un mot monosémique (*fonctionnaire*) et l'autre d'un mot polysémique (*commence*). Chacune des propositions d'analyse comporte les deux types d'information sémantique prévus, c'est-à-dire la catégorie (`humain`) et

le domaine (`DOM_ADM`) pour un nom, le domaine (`DOM_TPS`, `DOM_ENS` etc.) et la classe (`CLA_X4a`, `CLA_M4b` etc.) pour un verbe. Pour le nom monosémique, l'information sémantique a simplement été ajoutée à la proposition d'analyse. Pour le verbe polysémique, chaque proposition d'analyse a été reproduite autant de fois que ce verbe a de sens tout en recevant l'information sémantique d'un des sens.

4.3.2 Intégration d'une taxinomie sémantique hiérarchique

La mise en correspondance de requêtes avec les éléments textuels susceptibles d'y apporter une réponse n'est pas une opération triviale. Nous avons signalé déjà plusieurs techniques qui concourent à ce but, et il en existe d'autres. Une de ces techniques consiste à généraliser les unités lexicales qui constituent la requête en une forme sémantique plus ou moins abstraite et de faire de même avec les lexèmes qui forment les segments de texte candidats à y apporter une réponse [Voorhees, 1993, Vossen, 1997]. Dès lors, si les formes sémantiques abstraites des éléments constitutifs de la requête et du texte coïncident, il est probable qu'ils désignent une même réalité, ou du moins une réalité semblable. Ainsi, les mots *fille* et *enfant* ne correspondent pas sur le plan lexical, mais la généralisation de ces lexèmes au travers du domaine d'application du dictionnaire *Dubois* fournit l'information `PAR` pour *parenté*.

Une généralisation peut donc être réalisée au travers de l'information sémantique du *Dubois*, le domaine d'application pour l'ensemble du lexique, la classe sémantique pour la catégorie verbale uniquement. Toutefois, ces domaines correspondent à une **structure plate** et ne sont pas hiérarchisés entre eux. Par exemple, pour les domaines `ROM` (*antiquité romaine*) et `GRE` (*antiquité grecque*), aucun domaine `ANT` (*antiquité*) n'existe qui regroupe les deux autres. Il faut dès lors se contenter d'un seul niveau de généralisation dont la granularité est définie par la ressource utilisée.

Or certaines ressources lexicales que nous avons précédemment mentionnées disposent d'informations sémantiques composées en classes hiérarchisées, qui permettent de faire varier le niveau de généralisation en l'augmentant ou en le diminuant par une navigation verticale entre les différents niveaux hiérarchiques. Ces ressources sont *EuroWordNet* français et *AlethDic*. Les importantes lacunes, notamment verbales, et les particularités lexicales du dictionnaire *AlethDic* ont justifié précédemment son élimination, mais *EuroWordNet*, quoique son étendue lexicale soit restreinte, peut sur une grande partie du lexique fournir ses indications hiérachiques importantes, tant dans la taxinomie hypéronymique que dans la taxinomie méronymique.

L'intégration de la structure sémantique à notre système pose toutefois certains problèmes. Tout d'abord, les entrées lexicales de *EuroWordNet* ne

sont pas découpées en acceptions, mais la détermination de leurs différents sens n'a de réalité que dans la mesure où un même lexème peut appartenir à différents ensembles synonymiques (*synsets*), chacun de ces ensembles représentant une signification particulière du lexème. La hiérarchisation sémantique de *EuroWordNet* s'appuie d'ailleurs sur ces ensembles synonymiques qu'elle classifie et structure, plutôt que sur les unités lexicales. D'autre part, la limitation du lexique couvert par *EuroWordNet*, que nous avons signalée dans la section qui lui est consacrée (*cf.* section 3.4.3 page 120), ne peut à l'évidence autoriser une généralisation que dans le nombre de cas, forcément limité, où les unités lexicales traitées dans les énoncés appartiennent au vocabulaire de la ressource.

Le problème de la répartition des différentes acceptions pour chaque lexème représenté s'apparente aux difficultés que nous avons rencontrées lors de la distribution des synonymes aux différentes acceptions d'un même mot. Dans le cas présent, les incompatibilités toujours affichées dans la subdivision sémantique de deux ressources lexicales s'y ajoutent, car lors de la distribution des synonymes proposés par *EuroWordNet* pour un lexème donné, nous n'avons tenu aucun compte des ensembles synonymiques qui forment la structure sémantique interne du réseau sémantique. Nous avons en effet pris le parti de respecter les choix du dictionnaire *Dubois*, car c'est essentiellement sa structure qui nous permet d'atteindre des informations adaptées au contexte dès lors que la désambiguïsation sémantique est effectuée.

Dans le cas présent, nous proposons de choisir comme arbre taxinomique celui dans lequel ce lexème apparaît comme un nœud et où son nœud-mère présente avec le lexème les mêmes similitudes que celles que nous avons notées pour la distribution des synonymes. Les unités lexicales désignées par un nœud-mère doivent dès lors, si elles ne sont pas verbales, posséder le même domaine du *Dubois* que leur nœud-fille, et s'il s'agit d'unités verbales, elles doivent appartenir aux mêmes classe et sous-classe sémantiques que leur nœud-fille, ou posséder le même domaine et la même classe sémantique. Cette procédure est valable pour les deux types de taxinomies présents dans la ressource *EuroWordNet*.

Le temps nous a manqué pour réaliser ne serait-ce que le filtrage des arbres taxinomiques pour élargir à une structure hiérarchique l'information lexicale destinée à enrichir le texte. Dès lors, nous n'avons pas pu tester la validité de la méthode de choix des arbres taxinomiques, ni mettre en œuvre cette méthode. À plus forte raison nous n'avons pu réaliser la procédure de généralisation des termes que ce soit dans le corps des documents ou dans les requêtes proposées au système.

4.4 Conclusion

Face à l'information riche et variée, mais disparate voire contradictoire de plusieurs ressources lexicales, nous avons dû étudier les possibilités d'exploiter ces données tout en conservant un niveau de compatibilité acceptable entre elles. Nous avons dès lors mis en œuvre des méthodes automatiques qui permettent, à partir d'un dictionnaire, de verser dans un autre un type de données qui en est absent en s'appuyant sur des indications communes aux deux ressources.

Nous avons donc pu constituer des ensembles synonymiques plus importants sans perdre la spécificité de la synonymie propre à une acception plutôt qu'à une entrée. Nous avons également constitué des champs dérivationnels qui correspondent aux lemmes qui le permettent, pour lesquels nous avons également dû construire manuellement des schémas de correspondance syntaxique. Nous avons encore permis d'attribuer aux unités lexicales des traits et classes sémantiques et des domaines d'application lors de l'utilisation du lexique morphologique. Enfin, et sans avoir pu réaliser cette opération, nous avons étudié l'opportunité d'exploiter d'autres relations sémantiques à partir de dictionnaires adaptés.

Les modifications et corrections apportées aux ressources lexicales tant pour l'analyse textuelle que pour l'enrichissement des énoncés permettent maintenant d'aborder les traitements des documents avec des outils et des dictionnaires capables de subvenir aux besoins de la tâche. Le chapitre suivant s'y consacre.

Chapitre 5

Enrichissement des documents

5.1 Introduction

La création de la structure informationnelle est guidée par des besoins au confluent de deux mouvances : les disciplines d'extraction d'information et de question-réponse. La démarche choisie consiste à identifier dans un texte une information recherchée grâce aux différents aspects que cette information peut prendre. Toutefois, les formes sous lesquelles l'information peut se présenter ne sont pas produites au départ de la recherche (un type informationnel dans le cadre de l'extraction d'information, une requête dans une application de question-réponse), mais au départ de la base textuelle dans laquelle on recherche l'information. Ce choix a été motivé par la constatation que l'identification sémantique de données textuelles est facilitée par un contexte plus étendu. Ce contexte est généralement plus étendu dans le texte que dans un type informationnel ou dans une question.

La création d'une structure informationnelle permettant l'accès à l'information contenue dans une base textuelle repose sur deux principes : l'identification des éléments d'information contenus dans la base documentaire ainsi que des liens qui les unissent, et la production du plus grand nombre de formes différentes qui peuvent être prises par chaque donnée présente.

L'identification des éléments d'information est opérée par les différentes étapes d'analyse des documents : identification des lexèmes grâce à la segmentation et à l'analyse morphologique, identification des relations entre lexèmes et groupes de lexèmes au travers de l'analyse syntaxique, identification du sens des lexèmes grâce à la désambiguïsation sémantique. La collecte de ces résultats constitue l'épine dorsale de la structure informationnelle. L'ensemble de l'information textuelle que notre méthode est capable de recueillir y est présente.

Ensuite, la génération des expressions équivalentes aux formes originales doit respecter deux principes :

- c'est le sens de l'expression originale qui guide le choix des expressions générées ;
- l'expression générée doit s'intégrer dans la structure syntaxique de l'énoncé original en transformant son aspect sans en modifier exagérément le sens.

Les informations lexico-syntaxiques et lexico-sémantiques obtenues grâce aux ressources lexicales peuvent dès lors être ajoutées à la structure informationnelle pour l'enrichir.

Ce chapitre présente la méthode de constitution de la structure informationnelle et les techniques qui permettent d'obtenir les informations qui la constituent. Dans un premier temps, ce sont les résultats de l'analyse linguistique de la base textuelle qui en forment le squelette. Un nouveau système de réduction de l'ambiguïté sémantique mieux adapté aux besoins de la tâche est présenté. Ensuite, les traits sémantiques, les **synonymes simples** et les expressions synonymiques à mots multiples viennent enrichir la structure. Enfin, l'application de patrons syntaxiques permet d'ajouter les formes dérivées du lexique original et ainsi de compléter la structure informationnelle.

5.2 Extraction et stockage de l'information issue de l'analyse syntaxique

La structure de l'information contenue dans un texte correspond dans notre approche à l'ensemble des résultats que nous pouvons obtenir suite à une analyse linguistique de ce texte additionné de l'enrichissement que nous avons pu y apporter. L'analyse que nous effectuons repose sur un groupe d'outils – *NTM* et *XIP* – décrits dans la section 2.2 page 57. Cette analyse est fonction de la qualité du lexique morphologique et du niveau de compétences de la grammaire de *XIP*. Le lexique morphologique permet le découpage du document en unités lexicales. La grammaire effectue la désambiguïsation morpho-syntaxique des interprétations morphologiques proposées par *NTM*, construit un arbre d'analyse qui constitue les unités lexicales de chaque phrase en syntagmes minimaux et enfin établit les dépendances syntaxiques entre les lexèmes et entre les syntagmes.

L'information qui nous est proposée à l'issue de cette phase d'analyse est donc lexicale, puisque le découpage en mots est réalisé, elle est aussi morphologique à travers l'analyse de la forme des mots et leur désambiguïsation catégorielle, elle est enfin syntaxique grâce aux syntagmes minimaux et dépendances. Parmi ces différentes données, nous devons déterminer celles qui composent une part de la structure informationnelle du texte afin de la

stocker et de l'indexer pour y avoir accès lors d'une recherche d'information.

Ces données sont d'abord lexicales et le lemme de chaque unité lexicale doit impérativement être conservé en tant qu'entité porteuse de l'information sémantique de base dans le texte. D'un point de vue morphologique, il est important également de préserver la catégorie grammaticale de chaque lexème, car cette donnée est une indication importante pour toute la consultation ultérieure d'une ressource lexicale, et elle peut permettre de distinguer une interprétation parmi plusieurs propositions. Cette information morphologique ne doit être conservée que dans la mesure où la désambiguïsation catégorielle a déjà été effectuée.

Enfin, nous conservons principalement les données syntaxiques qui doivent servir de support aux relations syntaxico-sémantiques entre les concepts actualisés par les lexèmes. Dès lors, les syntagmes ne présentent pas un intérêt déterminant pour notre objectif contrairement aux dépendances qui sont porteuses de liens significatifs entre les concepts. Certaines de ces **dépendances** sont toutefois purement « **fonctionnelles** », et ne devront pas être stockées[1], tandis que d'autres sont plus significatives et doivent impérativement être conservées dans une perspective de relations syntaxico-sémantiques.

Ainsi, une dépendance `DET(detX, nomY)`, qui indique qu'un article `X` détermine une unité lexicale nominale `Y`, ne doit pas être considérée comme une relation importante pour notre application. *A contrario*, la relation `SUBJ(verbeA, nomB)`, qui indique que le nom `B` est sujet d'un verbe `A`, relie souvent un actant à une action, comme une dépendance `VARG[DIR](verbeA, nomC)`, pour indiquer que le nom `C` est le complément d'objet direct du verbe `A`, relie souvent l'action au **patient** qui la subit.

Cependant, nous avons décidé de ne pas trop préjuger de l'intérêt ou non de telle ou telle dépendance. Pour les relations de type sujet et objet direct dont nous venons de parler, leurs caractéristiques syntaxico-sémantiques ne dénotent pas systématiquement des qualités d'actant, d'action ou de patient. Le lexique peut faire varier ces opérateurs, ainsi que le reste de la structure syntaxique de la phrase. Dès lors, nous éliminons les seules dépendances qui concernent des mots purement grammaticaux[2], partant du principe que ces mots vides ne sont pas descriptifs d'entités contenues dans le texte [Martinet, 1960]. Nous conservons toutefois les dépendances pré-

[1] Nous avons choisi d'éliminer les dépendances fonctionnelles de la structure informationnelle car elle n'étaient pas utiles pour l'application que nous faisons de cette structure. Cette éviction est débrayable, de telle sorte qu'il est possible de maintenir ces dépendances dans la structure.

[2] On appelle **mots grammaticaux** (ou **mots-outils** ou **mots vides**) [Martinet, 1960, Grevisse et Goosse, 1991] les mots dont le rôle dans la phrase est plus grammatical que lexical. Leur nombre est limité et ils rassemble les pronoms, les conjonctions, les introducteurs, les prépositions, les pronoms, les déterminants, les auxiliaires et les copules.

positionnelles, c'est-à-dire les dépendances permettant de relier la tête du groupe prépositionnel à la tête du groupe dont dépend ce groupe prépositionnel (`NMOD[INDIR](X, prep, Y)`). En effet, même si c'est la préposition qui est fondatrice de cette dépendance, la relation unit en réalité la tête des autres syntagmes décrite par chacun des autres arguments de la dépendance (`X` et `Y`).

Grâce aux possibilités de la méthode de stockage de Claude Roux (*cf.* [Roux et Jacquemin, 2002] et annexe A page 275), chaque élément d'information est classé dans une base de données. Cet élément reçoit une indexation à différents niveaux de découpage du document : dépendance, phrase, paragraphe, texte. Cette classification à différents niveaux permet de définir, lors de la phase d'interrogation, une échelle d'exigence dans l'étendue de la fenêtre dans laquelle les éléments de la réponse communs avec ceux de la requête doivent être trouvés.

Le stockage de l'information obtenue au niveau de l'analyse syntaxique n'est pas cependant la seule opération que nous effectuons à cette étape de notre méthode. En effet, dès ce niveau, nous entrons dans un domaine proche de la sémantique et certaines distinctions entre des dépendances, pertinentes dans le cadre strictement syntaxique, ne le sont plus en ce qui concerne le sens de l'énoncé.

Toutefois, notre action à ce stade de la méthode ne se limite pas au seul stockage de l'information obtenue au travers de l'analyse syntaxique. En effet, dès ce niveau, il est possible de toucher à certains aspects plus sémantiques des résultats collectés. Il s'agit de mettre en correspondance certaines dépendances syntaxiques dont la distinction pour notre application n'est pas pertinente, afin de préserver une unité dans le sens plutôt qu'une distinction dans la structure syntaxique.

Ainsi, certaines relations syntaxiques construites par *XIP* et bien distinctes dans sa grammaire sont considérées comme équivalentes du point de vue du sens. Les dépendances syntaxiques différentes mais équivalentes que l'analyse syntaxique génère doivent donc être fusionnées sous une seule dénomination avant d'être stockées dans la base de données qui conserve l'information identifiée ou extraite du texte. Les équivalences de dépendances syntaxiques que nous avons constatées correspondent à une amélioration que nous avions réalisée sur le module de désambiguïsation sémantique développé à XRCE (Xerox Research Centre Europe), et qui avait été testé avec succès [Brun et al., 2001]. Elles ont toutefois été adaptées à la grammaire de *XIP* et sont exposées dans le tableau 5.1 page ci-contre.

Dans ce tableau, on peut voir que certaines constructions sont sémantiquement équivalentes à d'autres. Par exemple, la mise en correspondance de `SUBJ[PASS](X, Y)` et `VARG[DIR](X, Y)` permet d'inférer la conformité sé-

Dépendance syntaxique	Exemple	Dénomination équivalente
SUBJ [a]	Une partie des troupes se rallia à Élagabal. SUBJ(rallia, partie)	SUBJ(rallia, partie)
DEEPSUBJ [b]	Constantin fut proclamé auguste par les troupes de Bretagne. DEEPSUBJ(proclamé, troupes)	SUBJ(proclamé, troupes)
SUBJCLIT [c]	« Qu'ils me haïssent, pourvu qu'ils me craignent », disait-il. SUBJCLIT(disait, il)	SUBJ(disait, il)
VARG[DIR] [d]	César [...] écrase une armée des partisans de Pompée à Thapsus. VARG[DIR](écrase, armée)	VARG[DIR](écrase, armée)
SUBJ[PASS] [e]	Constantin fut proclamé auguste par les troupes de Bretagne. SUBJ[PASS](proclamé, Constantin)	VARG[DIR](proclamé, Constantin)
NMOD[NOUN, SPRED] [f]	Antoine fut l'ami et le second de César. NMOD[NOUN, SPRED](Antoine, ami)	NMOD[NOUN, SPRED](Antoine, ami)
SEQNP [g]	Constance III épousa Galla Placidia, sœur d'Honorius. SEQNP(Galla Placida, sœur)	NMOD[NOUN, SPRED](Galla Placida, sœur)

[a] Une dépendance SUBJ(X, Y) relie un verbe X à son sujet Y.

[b] Une dépendance DEEPSUBJ(X, Y) relie un verbe X et un mot Y considéré comme son sujet sémantique, soit le complément d'agent si X est à la voix passive, soit un infinitif complétif de X dans une construction impersonnelle.

[c] Une dépendance SUBJCLIT(X, Y) relie un verbe X à son sujet Y, Y étant un pronom personnel enclitique.

[d] Une dépendance VARG[DIR](X, Y) relie un verbe X à son objet direct Y.

[e] Une dépendance SUBJ[PASS](X, Y) relie un verbe X à la voix passive à son sujet Y.

[f] Une dépendance NMOD[NOUN, SPRED](X, Y) relie un nom X sujet d'un verbe copule et son attribut nominal Y.

[g] Une dépendance SEQNP(X, Y) relie un nom X et un nom Y qui lui est apposé.

TAB. 5.1 – Correspondances sémantiques de dépendances syntaxiques.

mantique entre les énoncés *Constantin fut proclamé auguste par les troupes de Bretagne* , où *proclamé* et *Constantin* sont unis par une dépendance SUBJ[PASS](proclamé, Constantin) et un énoncé *Les troupes de Bretagne proclamèrent Constantin auguste*, où les mêmes lemmes sont unis par une dépendance VARG[DIR](proclamèrent, Constantin). Il en va de même pour les autres dépendances mises en correspondance. Du fait de ces rassemblement de dépendances sous une seule dénomination, c'est un fragment de la sémantique de la phrase elle-même qui est emmagasinée au travers de ces relations syntaxiques.

5.3 Une nouvelle implémentation de la désambiguïsation sémantique

L'analyse morpho-syntaxique nous a donc permis d'identifier précisément les unités lexicales présentes dans le document, ainsi que les relations syntaxiques qui les relient. Toutefois, cette identification ponctuelle n'est pas apte à résoudre notre problème. Il s'agit à présent de dépasser ses limites par la détermination du sens des lexèmes dans leur contexte afin d'approcher la sémantique des documents et des unités qui les composent. C'est sur cette opération que repose la procédure d'enrichissement destinée à favoriser la mise en correspondance de la réponse avec sa requête. La méthode de désambiguïsation sémantique de XRCE, que nous avons décrite dans la section 2.3 page 75 [Dini et al., 1998, Segond et al., 1998], est appelée à remplir cette tâche d'identification de la signification des mots polysémiques en même temps qu'elle identifie des informations liées au sens des unités lexicales qui permettront d'effectuer un enrichissement du texte adapté à sa signification.

En effet, si nous avons choisi cette méthode de désambiguïsation sémantique pour identifier les concepts présents dans le texte, le système dans lequel cette méthodologie a été implanté n'est pas propre à une utilisation au sein de cette application. De fait, la plate-forme linguistique *XeLDA* ne dispose pas des analyseurs *NTM* et *XIP* que nous avons sélectionnés pour effectuer l'analyse morpho-syntaxique, ni des ressources lexicales qui nous seront utiles, à savoir *Dubois*, *Memodata*, *Bailly* et *EuroWordNet*, d'autant plus que nous avons amélioré ces ressources. De plus, la possibilité d'ajouter librement des règles à la grammaire de *XIP* va nous permettre de laisser la plus grande part de l'application des règles de désambiguïsation sémantique à l'analyseur syntaxique lui-même plutôt que de redéfinir toute une architecture d'application de ces règles sur le résultat fourni par l'analyse syntaxique. Du fait des besoins de notre méthodologie, les modifications à apporter au précédent système étaient si nombreuses et si profondes qu'une nouvelle implémentation s'est imposée, qui intègre de nouvelles fonctionnalités.

La méthodologie sur laquelle nous avons résolu de nous baser comporte deux étapes. Tout d'abord l'extraction de l'information lexicale permettant de définir les contextes d'une acception afin d'en déduire les règles de sélection de sens. Ensuite l'application de ces règles de sélection pour déterminer le sens des lexèmes polysémiques en fonction de leur contexte d'apparition. Ces deux étapes sont maintenues dans la nouvelle implantation que nous avons réalisée de la méthode de XRCE, mais certaines différences méritent d'être signalées.

5.3.1 Génération des règles de désambiguïsation

Le système de désambiguïsation sémantique de XRCE proposait un type de règles conditionnelles dont la condition s'appuyait sur le contexte du mot à désambiguïser[3]. Le succès de cette condition lors de l'application de la règle permettait de proposer un numéro de sens à ce mot et le système se chargeait d'afficher le fragment de l'article du dictionnaire correspondant à ce sens. Nous avons maintenu ce principe de règle conditionnelle sans toutefois prévoir un affichage du sens correspondant. Nous verrons la raison de cette différence dans le paragraphe chargé de décrire l'application des règles de désambiguïsation sémantique.

La méthodologie originale fonde la partie conditionnelle de chaque règle sur l'information que le dictionnaire contient sur le contexte de chaque mot dans chacune de ses acceptions possibles. Dans notre système, l'énonciation conditionnelle du contexte linguistique de la cible dans chacun de ses sens est également fonction de l'information contextuelle disponible dans le dictionnaire qui sert de référence lexico-sémantique à cette tâche. Dans le cas du dictionnaire *Dubois*, le contexte linguistique de chaque lemme dans chacune de ses acceptions peut être exprimé de trois manières.

Les règles lexicales

La première des informations qui permettent d'identifier le contexte d'une expression à désambiguïser est lexico-sémantique. Elle correspond aux règles d'exemples du précédent système et consiste à établir une règle de désambiguïsation sémantique à partir de l'analyse syntaxique des exemples fournis par le dictionnaire. Chacune des dépendances extraites de l'exemple par l'analyse de *NTM-XIP*, et qui implique la cible, dénote en effet un contexte linguistique propre à ce mot dans le sens considéré. Chacune des dépendances extraites dont un des arguments est la cible est dès lors susceptible de constituer la condition contextuelle d'une règle de désambiguïsation permettant de discriminer le sens considéré de ce mot.

Cependant, nous avons signalé lors de la description du dictionnaire *Dubois* (*cf.* section 3.2 page 100) que la présentation du champ informationnel de l'exemple se prêtait peu à un traitement automatique du fait de l'abréviation (l'initiale suivie d'un tilde représente une forme du lemme) de l'entrée dans son énoncé. Ainsi, l'exemple de la neuvième acception du verbe *admettre*[4] se présente sous la forme :

Ce texte a∼ une seule interprétation.[5]

[3] Nous désignons sous le nom de « cible » ce mot à désambiguïser.
[4] De sens « supporter », « accepter comme valable ».
[5] *Ce texte* ***admet*** *une seule interprétation.*

Étant donné qu'une forme comme *a∼* ne peut être identifiée correctement, l'analyse de cet énoncé sera généralement erronée. La figure 5.1 montre que cet exemple ne peut être analysé tel quel : la forme *a∼* y est considérée comme un nom apposé au lexème *texte* (`NN`), et les seules autres dépendances construites relient soit *seule* à *interprétation* comme épithète (`NMOD[ADJ]`), soit les déterminants *Ce* et *une* au nom qu'ils déterminent (`DETERM`), respectivement *texte* et *interprétation*. Cette analyse est bien entendu incorrecte. Il s'agit donc de signaler à l'analyseur que la forme abrégée présente dans chaque exemple est un lexème qui appartient à la catégorie lexicale de l'entrée dans laquelle apparaît l'exemple. *XIP* dispose ainsi des éléments nécessaires pour effectuer au mieux son analyse.

```
$> echo "Ce texte a∼ une seule interprétation." | xip

NMOD[ADJ](interprétation,seule)
NN(texte,a∼)
DETERM(Ce,texte)
DETERM(une,interprétation)

0>GROUPE{NP{Ce texte} NP{a∼} NP{une AP{seule}
interprétation} .}
```

FIG. 5.1 – Analyse problématique d'un exemple contenant une forme abrégée.

Une particularité de *XIP* est de permettre la création de règles lexicales capables de donner à un mot ou à une chaîne de caractère des traits quelconques, y compris une catégorie grammaticale. Lors de l'analyse du dictionnaire, nous sommes donc en mesure de générer pour chaque entrée une règle lexicale dans le formalisme *XIP* qui impose à toute unité lexicale se présentant sous la forme de l'initiale de la vedette suivie d'un tilde (ici : *a∼*) un trait lui conférant la catégorie grammaticale de cette même entrée. Par la suite, la forme abrégée est interchangée avec le lemme du mot-vedette lors de la construction de la règle de désambiguïsation. La figure 5.2 page ci-contre présente un exemple de règle lexicale imposant la catégorie verbale à toute chaîne de caractères *a∼*, puis le résultat de l'analyse de l'énoncé provenant du champ d'exemple du *Dubois*, qui présente correctement *texte* comme sujet de *a∼*, ainsi que *interprétation* comme complément d'objet direct de *a∼*.

Toutefois, certaines relations extraites d'un exemple ne doivent pas être considérées comme pertinentes pour la création de règles pour la sélection du sens des mots en contexte. En effet, dès lors que ces dépendances sont purement fonctionnelles[6], ou qu'elles font intervenir des mots grammaticaux,

[6]Nous désignons par le terme « dépendances fonctionnelles » les dépendances que *XIP* extrait pour son fonctionnement interne, et qui prennent généralement comme argument

Règle lexicale :
a∼ = Verb

```
$> echo "Ce texte a∼ une seule interprétation." | xip

SUBJ(a∼,texte)
VARG[DIR](a∼,interprétation)
NMOD[ADJ](interprétation,seule)
DETERM(Ce,texte)
DETERM(une,interprétation)

0>GROUPE{SC{NP{Ce texte} FV{a∼}} NP{une AP{seule}
interprétation} .}
```

FIG. 5.2 – Analyse correcte d'un exemple par résolution de la forme abrégée.

leur caractère devient trop général et trop commun pour réaliser une distinction pertinente entre les différentes acceptions d'un même lexème. Ces dépendances sont donc éliminées avant la construction des règles de désambiguïsation sémantique, exclusivement axées sur un lexique et des relations significatives.

Pour illustrer le mode de construction des règles lexicales, nous reprenons l'exemple extrait du neuvième sens du verbe *admettre* dans le dictionnaire *Dubois* (figures 5.1 page précédente et 5.2). Cet exemple est présenté à la chaîne d'analyse *NTM-XIP* après que la règle lexicale correspondant à sa forme abrégée a été ajoutée dans la grammaire de *XIP*. Parmi les dépendances extraites, seules sont conservées celles qui impliquent le mot-vedette [7], à condition que ce ne soient pas des relations fonctionnelles. À partir de ces dépendances, la condition de la règle de désambiguïsation est construite dans le formalisme de *XIP*, `#0` correspondant au lemme du mot-vedette tandis que les autres arguments des dépendances sont représentés par les autres variables en `#`.

---

des unités lexicales appartenant à un même syntagme. Ces dépendances sont :
- `AUXIL` entre un auxiliaire conjugué et sa base verbale ;
- `CLOSEDNP` qui porte sur un nom apte à être la tête d'un syntagme nominal ;
- `DETERM` entre un déterminant et le nom qu'il détermine ;
- `MWEHEAD` entre une expression composée de plusieurs mots et le mot de cette expression qui en est la tête. C'est l'expression à mots multiples qui est utilisée comme argument d'autres dépendances ;
- `PREPOBJ` entre une préposition et la tête du groupe prépositionnel. Cette dépendance est utilisée pour effectuer ultérieurement le rattachement prépositionnel ;
- `PRECOMMA` qui porte sur la tête du premier syntagme qui suit une virgule ;
- `PUNCT`

Elles n'entrent pas dans la construction des règles de désambiguïsation sémantique.

[7] Elles sont présentées en caractères **gras** à la figure 5.3 page suivante.

```
$> echo "Ce texte a ~ une seule interprétation." | xip

SUBJ(a ~, texte)
VARG[DIR](a ~, interprétation)
NMOD[ADJ](interprétation, seule)
DETERM(Ce, texte)
DETERM(une, interprétation)
```

FIG. 5.3 – Construction d'une règle lexicale de désambiguïsation.

Règles de sous-catégorisation

L'information syntaxico-sémantique est la deuxième à être exploitée pour déterminer le contexte typique d'un mot dans un sens donné. Il s'agit des données de sous-catégorisation fournies par le dictionnaire *Dubois*, qui précisent le ou les schémas syntaxiques propres à l'unité lexicale dans chacun de ses sens et fournissent le cas échéant un trait sémantique tenant à la nature de l'argument des dépendances (*humain*, *animal*, *inanimé*), sans que l'unité lexicale qui correspond à cet argument soit précisée.

Lors de l'examen des différents champs informationnels du dictionnaire, nous avons toutefois signalé certaines carences et certaines imprécisions dans le champ informationnel de sous-catégorisation. Principalement, nous n'avons pas pu résoudre le problème lié à la catégorisation prépositionnelle des compléments circonstanciels, et de ce fait, nous ne pouvons exploiter les directives qui portent sur la nature de la préposition introduisant ces groupes pour décider du sens à donner à un mot dans son contexte. Pour éviter des erreurs par trop évidentes, nous n'utilisons donc pas cette information prépositionnelle des circonstants dont nous ne pouvons être sûr dans une marge acceptable.

Comme nous le disions précédemment, les informations de sous-catégorisation dont la qualité est suffisamment satisfaisante pour que nous l'exploitions se présentent sous la forme de schémas syntaxico-sémantiques, c'est-à-dire que certaines relations syntaxiques qui impliquent le lemme de l'entrée sont obligatoires, possibles ou interdites selon les cas, et que des contraintes sémantiques peuvent peser sur les arguments de ces dépendances lorsqu'elles ne sont pas interdites.

Ainsi, des schémas syntaxiques de sous-catégorisation existent traditionnellement dans les ressources lexicales, avec les notions de pronominalité, de transitivité ou d'intransitivité d'un verbe, dénotant respectivement l'obligation d'une relation syntaxique entre le verbe et un pronom personnel réfléchi, la possibilité d'une relation sémantique – directe ou indirecte selon les cas

– entre le verbe et son objet ou l'interdiction d'une relation directe entre le verbe et un objet. Outre ces informations syntaxiques fréquemment présentes dans les dictionnaires, d'autres schémas relationnels sont prescrits dans le *Dubois*. De plus, des contraintes sémantiques pèsent souvent sur les arguments unis par les relations syntaxiques à l'unité lexicale correspondant au lemme de l'entrée. Ces arguments doivent selon les cas présenter un trait *humain*, *animal* ou *inanimé*.

Pour former des règles de désambiguïsation sémantique, ces schémas de sous-catégorisation doivent être traduits en conditions d'application dans le formalisme de *XIP*. Ici encore, les fonctionnalités de cet outil d'analyse permettent de construire aisément ces règles. En effet, les dépendances syntaxiques de la grammaire du français peuvent facilement être mises en correspondance avec les relations syntaxiques prescrites par la ressource lexicale, et les contraintes sémantiques sur les arguments des relations peuvent être intégrées sous forme de traits obligatoires.

On se souviendra que les traits sémantiques qui correspondent aux contraintes sémantiques des schémas de sous-catégorisation ont été intégrés au lexique lors de la phase d'élargissement de l'information du lexique utilisé par *NTM* pour effectuer le découpage du texte et son analyse morphologique (*cf.* section 4.3.1 page 145). De cette manière, ces traits sont automatiquement liés au vocabulaire utilisé dans les document lors de l'analyse linguistique de bas niveau. Ils sont donc accessibles dès l'application de la grammaire syntaxique et, à plus forte raison, lors de l'application des règles de désambiguïsation sémantique.

L'exemple 5.4 page suivante permet d'illustrer la technique de construction des règles de désambiguïsation sémantique à partir d'une information de sous-catégorisation. Le schéma syntaxique fourni par le dictionnaire *Dubois* impose à chaque sens de *falloir* une construction grammaticale différente. Chacun de ces sens possède en effet un schéma syntaxique propre et présente une catégorisation sémantique des arguments de certaines relations syntaxiques qui peuvent impliquer le mot-vedette. La mutation des différentes relations syntaxiques des schémas proposés en dépendances *XIP* et l'adaptation des contraintes sémantiques sur les arguments de ces relations en traits placés sur les arguments des dépendances permet de poser une ou plusieurs conditions au choix d'un sens, d'un domaine et pour les verbes d'une classe sémantique. La condition de chaque règle peut donc porter sur la nature d'une ou plusieurs dépendances et sur un trait sémantique attaché à un argument de dépendance autre que le mot-vedette. Dans les cas de transitivité indirecte ou d'intransitivité, il est possible de refuser l'application de la règle si une relation d'objet direct apparaît, traduite sous la forme `VARG[DIR]`. Une restriction peut également apparaître sur le lexique, comme on peut le voir dans la définition de la préposition *à* qui introduit le

Entrée	Sens	Schéma
falloir 01	besoin, convenance	T3500
falloir 02	nécessité	N4a A40
falloir 03	éloignement, manque	P4000

Construction des règles correspondant à chacun des schémas :

T3500 –> verbe transitif, sujet (inanimé), complément direct

```
falloir : Verb =
    if SUBJ(#0,#1[inanimé :+])
           & VARG[DIR :+](#0, #2)
    & #0[GRA=+,X2a=+] & #0[n1=+]
```

N4a –> verbe transitif indirect, sujet (inanimé/complétive), complément indirect en *à*

```
falloir : Verb =
    if ( SUBJ(#0,#1[inanimé :+])
             || COMPLETIVE(#0,#2)
             || INFINITIVE(#0,#3) )
           & ~VARG[DIR :+](#0, #4)
           & VARG[INDIR :+](#0, #5[lemme=à], #6)
    & #0[GRA=+,X1a=+] & #0[n2=+]
```

A40 –> verbe intransitif, sujet (inanimé/complétive)

```
falloir : Verb =
    if ( SUBJ(#0,#1[inanimé :+])
             || COMPLETIVE(#0,#2) )
           & ~VARG[DIR :+](#0, #3)
    & #0[GRA=+,X1a=+] & #0[n2=+]
```

P4000 –> verbe pronominal, sujet (inanimé/complétive)

```
falloir : Verb =
    if SUBJ[PRON](#0,#1[inanimé :+])
    & #0[QUA=+,X2a=+])&(#0[n3=+])
```

FIG. 5.4 – Exemple d'extraction de règle de sous-catégorisation.

complément indirect.

Les règles de domaine

Le troisième des différents types de règles de désambiguïsation sémantique correspond aux règles que le système de XRCE appelait « règles sémantiques » (*cf.* section 2.3.3 page 84). De même que dans cette approche que nous avons choisie comme point de départ pour la gestion du sens dans notre méthodologie, nous constatons que les règles d'exemples sont très contraintes

en même temps qu'elles sont extrêmement limitées dans leurs possibilités de s'appliquer, à moins que la cible se présente dans un contexte identique tant au niveau syntaxique que lexical. Il s'agit dès lors d'élargir les possibilités de couverture de ces règles.

À la suite des cognitivistes de l'intelligence artificielle [Masterman, 1961, Quillian, 1968] et des spécialistes intéressés par le traitement de l'information [Voorhees, 1993], le précédent système de désambiguïsation s'est basé sur le principe de l'interchangeabilité relative d'unités lexicales appartenant à la même catégorie lexico-sémantique. Ces catégories correspondaient tantôt aux classes sémantiques du *WordNet* anglais, tantôt à la hiérarchie de *AlethDic* pour le français. L'utilisation de catégories à la place de lexèmes permettait de généraliser chaque règle de désambiguïsation lexicale.

Tout en conservant cette approche généralisatrice, nous avons fait le choix dans la présente application de privilégier l'exploitation du domaine plutôt que celle de la classe sémantique, partant de la constatation très tôt faite en traduction automatique que le choix d'un domaine précis restreint l'ambiguïté des mots, et donc facilite la désambiguïsation [Weaver, 1949, Gale et al., 1992]. C'est donc une catégorie lexicale dépendante du domaine qui sert ici à élargir le champ d'application des règles d'exemple. Nous n'avons pas l'intention toutefois d'abandonner l'élargissement par classe de mots, mais le dictionnaire *Dubois* ne possède pas de taxinomie et l'intégration de celle de *EuroWordNet* dans le programme de génération des règles demandait un travail conséquent que nous n'avons pu mener à bien dans des délais raisonnables. Toutefois, les résultats apportés par ce type de règles sont d'un niveau plus élevé que ceux des règles sémantiques du précédent système de désambiguïsation [Jacquemin et al., 2002].

Nous construisons les règles de domaine directement à partir des règles de désambiguïsation lexicales. Ces règles, tirées de l'analyse syntaxique des exemples du *Dubois*, présentent en effet des contraintes sur les unités lexicales contenues dans les exemples et sur les relations syntaxiques qu'elles entretiennent entre elles. La génération des règles de domaine, plus générales, consiste à modifier les contraintes pour qu'elles portent non plus sur les unités lexicales, mais sur les domaines d'application que ces unités possèdent dans le dictionnaire *Dubois*.

Ainsi, dans la règle lexicale de notre exemple (*cf.* figure 5.3 page 160), les contraintes d'application portent sur deux unités lexicales : *texte* et *interprétation*. Or ces lexèmes possèdent respectivement les domaines `LIT` (pour littérature) et `LOQ` (pour *parole*), `SPE` (pour *spectacle*), `PSY` (pour *psychologie*) dans le dictionnaire *Dubois*. D'autre part, les domaines d'application sont des traits sémantiques affectés à toute unité lexicale dès lors qu'une analyse de *NTM* est mise en œuvre. Les contraintes sur les lexèmes sont donc remplacées, dans les règles de domaine, par des contraintes sur les traits de

domaine assignés aux lexèmes.

```
$> echo "Ce texte a~ une seule interprétation." |
xip
```

Règle lexicale de désambiguïsation :

```
admettre : Verb =
     if(SUBJ(#0,#1[lemme : texte]))
          &VARG[DIR](#0,#2[lemme : interprétation])
     &(#0[LIT=+,S4h=+]&(#0[n9=+]))
```

Domaines correspondant aux unités lexicales :

texte :	LIT
Interprétation :	LOQ
	SPE
	PSY

Règle de domaine :

```
admettre : Verb =
     if(SUBJ(#0,#1[LIT :+]))
          &VARG[DIR](#0,#2[LOQ :+ || SPE :+ || PSY :+])
     &(#0[LIT=+,S4h=+]&(#0[n9=+]))
```

FIG. 5.5 – Construction d'une règle de domaine pour la désambiguïsation sémantique.

La figure 5.5 illustre la méthode de construction des règles de domaine. Les domaines d'application des lemmes qui apparaissent dans la condition de chaque règle lexicale sont extraits du *Dubois*. Ensuite, les contraintes sur les unités lexicales sont remplacées par des contraintes sur les traits de domaine. Aucune condition n'apparaît plus sur la nature des unités lexicales. On peut constater que lorsqu'un des lexèmes de la règle lexicale possède plusieurs domaines, chacun de ces domaines est susceptible de permettre l'application de la règle.

Les règles non verbales

Ces trois types de règles sont ceux que nous avons pu réaliser à partir de l'information contenue seulement dans la partie verbale du dictionnaire *Dubois*. La partie généraliste de ce dictionnaire ne comporte en effet ni exemple,

ni indication de sous-catégorisation. Il nous a donc fallu recourir à d'autres données pour pouvoir effectuer la désambiguïsation sémantique des lexèmes des catégorie grammaticale autres que verbales.

Les catégorie grammaticale qui nécessitent surtout une désambiguïsation du sens, à part les verbes, sont les noms et les adjectifs. Or le dictionnaire *Dubois* ne fournit, pour ces entrées, que peu d'information qui permette de distinguer leurs différentes acceptions. Il s'agit de la catégorie grammaticale elle-même, qui permet d'inhiber certaines propositions [Kelly et Stone, 1975], ainsi que des traits sémantiques (domaines d'application, catégories sémantiques pour les noms). Ces données sont donc seules en mesures de permettre une discrimination des sens des lexèmes nominaux et adjectivaux.

Dès lors, c'est sur base des catégorie grammaticale, des domaines d'application et des catégories sémantiques que les contraintes s'établissent pour la génération des règles de désambiguïsation sémantique des noms et des adjectifs. L'exploitation de la catégorie grammaticale ne pose aucun problème dès lors qu'une analyse morpho-syntaxique décide de la catégorie grammaticale de chaque unité lexicale présente dans un énoncé. Par contre, l'utilisation de contraintes sur les traits sémantiques n'est envisageable que dans la mesure où certains des traits sont activés et d'autres inhibés.

Les règles verbales de désambiguïsation sémantique font appel tantôt aux catégories sémantiques du contexte de la cible, lorsque ce sont des **règles de sous-catégorisation**, tantôt aux domaines d'application, lorsque le système fait appel à des règles de domaine. Il s'agit dès lors de donner un indice aux traits sémantiques utilisés par les règles verbales pour indiquer que ces traits ont été activés par la désambiguïsation sémantique. Les traits ainsi activés sur certaines unités lexicales dans le texte permettent de discriminer leurs acceptions.

Dans le corpus que nous avons constitué pour effectuer notre évaluation (*cf.* 7.2.3 page 203), près de 40 % des substantifs ont une dépendance syntaxique en commun avec un verbe. De plus, plus de 45 % des noms qui n'ont pas de relation avec un verbe ont une dépendance commune avec un des noms reliés à un verbe. Par ailleurs, plus de 90 % des adjectifs sont soit épithète, soit attribut d'un nom. La plupart des autres adjectifs qualifient un pronom.

Les règles de désambiguïsation sémantiques pour les adjectifs et les noms consistent donc dans un premier temps à utiliser les indices laissés par la désambiguïsation sémantique des verbes sur les traits de leur contexte syntaxique. Ensuite, les traits activés se propagent aux autres noms et adjectifs à travers les relations syntaxiques qu'ils entretiennent avec les premières unités lexicales désambiguïsées. Nous verrons à la section suivante, consacrée à l'application des règles de désambiguïsation sémantique, de quelle manière

ces règles s'agencent et fonctionnent.

générique 01	LIN	qui indique le type	adjectif/nom masculin inanimé
générique 02	**CIN**	**indications initiales**	**nom masculin inanimé**
générique 03	PHA	produit dans le domaine public	adjectif

Règle nominale :

```
générique : Noun =
     if (?(#1[verb :+],#0[CIN :+,inanime :+,desamb :+])
     || (~?(#1[verb :+],#0)
     & (?(#2[CIN :+,inanime :+,desamb :+],#0)
          ||?(#0,#3[CIN :+,inanime :+,desamb :+]))))
```

FIG. 5.6 – Construction d'une règle nominale de désambiguïsation sémantique.

La figure 5.6 illustre la construction d'une règle de désambiguïsation sémantique pour la seconde entrée de *générique*. Deux de ses acceptions sont nominales, et deux peuvent être adjectivales. La désambiguïsation catégorielle ne suffit donc pas à établir le sens correct. La règle de désambiguïsation se borne donc à vérifier la présence de traits sémantiques communs (`CIN` pour *cinéma*, et `inanimé`) sur la cible elle-même si elle a une dépendance syntaxique commune avec un verbe, quelle qu'elle soit (le `?` indique que la dépendance n'est pas déterminée), ou sur un nom ou un adjectif avec lequel la cible entretient une relation syntaxique, si cette cible n'a pas de relation avec un verbe. Dans tous les cas, le trait `desamb :+` est chargé de vérifier que l'unité qui sert à effectuer la désambiguïsation a elle-même été désambiguïsée.

Enfin, nous avons dû créer certaines « règles » d'assignation de sens pour les unités lexicales qui ne présentent pas d'ambiguïté sémantique. En effet, seuls les lexèmes qui comportent un numéro de sens pourront recevoir un enrichissement. Par ailleurs, ces unités lexicale n'étant pas ambiguës, elles peuvent être considérées comme désambiguïsées et servir dès lors de base à la désambiguïsation sémantique d'autres unités lexicales. La section suivante indique comment cette désambiguïsation peut s'effectuer.

5.3.2 L'application des règles de désambiguïsation sémantique

L'ensemble des règles dont nous avons décrit les différents modes de construction exploitent une information extraite du dictionnaire *Dubois*. Chaque règle concourt au choix d'un sens unique du lexème pour lequel elle a été élaborée, qui correspond au contexte dans lequel ce lexème apparaît. Cependant, la désambiguïsation sémantique n'est qu'une étape dans une approche plus large et l'application des règles peut dès lors varier en fonction des besoins de cette approche.

En effet, la phase de désambiguïsation sémantique s'inscrit ici dans un processus qui vise à construire une structure informationnelle qui correspond au contenu d'une base textuelle définie. Pour favoriser la manipulation des données contenues dans la base textuelle, cette structure doit être enrichie autant qu'il est possible de manière à présenter la plus grande variation de présentations sans que l'information qu'elle renferme ne soit modifiée. Tout enrichissement d'un énoncé étant susceptible de transformer le sens de cet énoncé, il s'agit en tout cas d'aboutir au meilleur compromis entre enrichissement maximal et sens inaltéré.

Dans notre méthodologie, toute la phase d'enrichissement est soumise à l'application de la désambiguïsation sémantique. Notre approche vise en effet une soumission de chaque information ajoutée au sens que possèdent les mots dans la phrase.

Cependant, la richesse de la structure sémantique est primordiale pour la gestion de l'information qu'elle contient. En effet, cette richesse seule permet d'atteindre une information souhaitée, quelle que soit la formulation de la requête. Par contre, la combinaison de plusieurs données dans une même requête constitue un filtre pour les informations enrichies fautivement dans la mesure où plusieurs de ces informations erronées correspondant à la requête n'apparaissent pas dans la même fenêtre de réponse.

Avec une précision de 74 % [Jacquemin et al., 2002], les résultats de cette désambiguïsation sémantique améliorent le potentiel de la méthode, mais ils ne sont pas suffisants car dans 26 % des cas, la désambiguïsation sémantique est fausse et provoque la perte de qualité de tous les enrichissements qui en découlent. Dès lors, notre position est de sélectionner parmi les sens proposés celui ou ceux qui sont les plus vraisemblables en fonction du contexte, au risque de sélectionner, à côté du sens exact, une ou plusieurs interprétations fausses, plutôt que de choisir un seul sens et de perdre le bon dans plus du quart des cas.

Les règles de désambiguïsation sémantique correspondent à un formalisme propre à *XIP*. Comme elles demandent une analyse morpho-syntaxique

pour s'appliquer, c'est au cours du fonctionnement de *XIP* que la désambiguïsation sémantique est mise en œuvre, dès que l'analyse syntaxique est terminée.

Avant toute désambiguïsation réelle, ce sont les règles liées aux lexèmes monosémiques qui doivent s'appliquer. Elles assignent un numéro de sens *1* à tous les mots qui ne présentent pas d'ambiguïté sémantique, et leur assignent un trait dénotant que la désambiguïsation a été effectuée pour eux. Nous verrons lors de l'application des règles non verbales l'intérêt de ce trait.

Les règles verbales, qui sont les seules règles « polysémiques » à s'appliquer directement sur les résultats de l'extraction des dépendances, sont les premières présentées à l'analyseur. L'exploitation des règles portant sur les autres catégorie grammaticale se fait ultérieurement.

Comme pour la méthode initiale, l'application des règles de désambiguïsation sémantique est soumise au respect d'une condition syntaxique et lexicale ou sémantique. La satisfaction de cette condition déclenche l'assignation d'une acception à la cible sous la forme d'un trait correspondant à son numéro de sens dans le dictionnaire.

Cependant, nous nous éloignons ici de la méthodologie appliquée à XRCE sur différents points. Tout d'abord, l'application des règles de sous-catégorisation et de domaine passe par la correspondance sémantique du contexte de la cible avec les traits sémantiques du schéma syntaxico-sémantique qui forme la condition de ces règles. Or notre système active par un marquage les traits sémantiques qui ont été utilisés lors de l'application d'une règle de désambiguïsation, afin de pouvoir effectuer ultérieurement la désambiguïsation d'autres unités lexicales. Le système original ne tenait pas compte des unités lexicales désambiguïsées pour effectuer son travail.

Ensuite, alors que l'application d'une règle de désambiguïsation faisait précédemment d'un sens de la cible un candidat privilégié pour devenir le sens unique de cette cible, chaque déclenchement d'une règle assigne maintenant directement le sens qui lui est propre à la cible, privilégiant en cela l'enrichissement et non la précision de la désambiguïsation sémantique.

Enfin, l'application des règles non verbales est postérieure à celle des règles verbales. En effet, ces règles non verbales dépendent de l'activation de traits sémantiques dont nous avons parlé précédemment. L'application des règles non verbales répond de fait à la satisfaction d'une condition syntaxico-sémantique dans laquelle les contraintes syntaxiques se résument à demander à une dépendance quelconque de mettre la cible en relation avec un lexème déjà désambiguïsé [8].

[8] De là l'intérêt de signaler les unités lexicale monosémiques comme désambiguïsées.

Énoncé à désambiguïser :
Il **fallait** un peu d'inconscience à François Hinard et à ses collaborateurs pour s'y risquer.

Dépendances impliquant *falloir* produites par *XIP* :
```
SUBJ[IMPERSO](fallait,Il[humain, animal, inanimé])
VARG[INDIR](fallait,François Hinard)
VARG[DIR](fallait,peu)
```

Règle de désambiguïsation correspondante :
```
falloir : Verb =
    if SUBJ(#0,#1[inanimé :+])
        & VARG[DIR :+](#0, #2)
    & #0[GRA=+,X2a=+] & #0[n1=+]
```

Résultat de la désambiguïsation :
falloir[GRA :+,X2a :+n1 :+] : falloir01 besoin, convenance.

FIG. 5.7 – Application d'une règle de désambiguïsation sémantique verbale (sous-catégorisation).

Si ce lexème est un verbe, les traits sémantiques correspondant à un sens au moins de la cible ont été activés et permettent d'assigner ce sens à la cible. Les liens syntaxiques qui unissent la cible et le verbe ont de plus été spécifiés par la règle verbale qui s'est appliquée et a activé certains traits sémantiques. Dans le cas où aucune dépendance ne relie la cible à un verbe désambiguïsé, la règle ne fonctionnera que dans la mesure où la cible est en relation syntaxique avec une unité lexicale désambiguïsée qui présente des traits communs avec un de ses sens. Le sens correspondant à ces traits sémantiques sera donc sélectionné. Dans ce dernier cas, aucune spécification particulière ne définit la dépendance syntaxique qui relie la cible et l'unité lexicale désambiguïsée.

Pour terminer, nous avons ajouté à l'application des règles une fonctionnalité en rapport direct avec le but poursuivi par notre démarche, à savoir un enrichissement maximal des documents analysés. Cette fonctionnalité n'est pas obligatoire. Après avoir constaté un rappel de 44 %, nous avons conclu que dans des cas trop nombreux, aucun enrichissement ne pourrait avoir lieu par défaut de sens choisi. Nous avons donc implanté la possibilité d'attribuer à toutes les unités lexicales non désambiguïsées l'ensemble des sens que le dictionnaire recense. Dès lors, les informations erronées ajoutées seront nombreuses pour ces unités lexicales, mais un enrichissement exact sera présent également.

5.4 Adjonction des synonymes

Du fait de l'action des règles de désambiguïsation sémantique, certains traits sémantiques relatifs aux domaines, classes et catégories ont été maintenus sur les unités lexicales qui composent la base textuelle, tandis que d'autres ont été inhibés. La présence de ces traits constitue le premier enrichissement des textes de la base documentaire.

De plus, un ou des numéros de sens ont été ajoutés sur chaque nœud lexical des arbres d'analyse syntaxique partiels lors de la désambiguïsation sémantique. Bien que ces traits ne constituent pas à proprement parler un enrichissement important, ils sont à la base de l'accès à une information lexicale adaptée au sens des mots présents dans les documents. En effet, toutes les données qui vont contribuer à l'enrichissement ultérieur de la base documentaire sont rendues accessibles au travers des numéros de sens, et chaque enrichissement peut être assigné comme trait au mot qu'il enrichit grâce à l'attribut `$STACK` dont la valeur est justement l'enrichissement (*cf.* section 2.2.2 page 64).

Le premier type d'enrichissement à réaliser à partir de ce résultat concerne l'apport de synonymes des unités lexicales au sens identifié. Grâce aux ressources lexicales que nous avons choisies (*cf.* chapitre 3 page 99) ou que nous avons adaptées (*cf.* chapitre 4 page 127), nous disposons d'ensembles synonymiques liés au sens des unités lexicales, et non plus à leur lemme. Nous pouvons donc aisément adjoindre aux lexèmes présents dans les documents les synonymes qui leur sont propres par simple consultation de leur numéro de sens.

Cependant, pour intégrer un synonyme dans les textes de la base documentaire, deux cas de figure peuvent se présenter : le synonyme est une seule unité lexicale, même composée ; ou bien il s'agit d'une expression formée de plusieurs lexèmes[9]. Le traitement pour intégrer une expression synonymique dans une phrase sera différent de celui qui y insère un synonyme simple.

5.4.1 Enrichissement par synonymes simples

Dans un premier temps, nous avons développé une procédure élémentaire pour intégrer un synonyme simple à l'énoncé. Cette procédure consistait à dupliquer chaque dépendance impliquant le mot à enrichir pour ensuite remplacer ce mot par son synonyme dans chacune des dépendances ainsi générées. Chacune de ces nouvelles dépendances est alors indexée dans la struc-

[9] Nous faisons la distinction entre « expression synonymique », c'est-à-dire un synonyme composé de plusieurs unités lexicales dont une au moins n'appartient pas à la même catégorie grammaticale que les autres, et « synonyme simple », qui correspond à une seule unité lexicale, même s'il peut s'agir d'un mot composé.

ture syntaxico-sémantique et chaque enrichissement peut ainsi permettre de retrouver l'information originelle dans les documents.

Énoncé à enrichir :
« Son règne a **favorisé** la décadence de la vieille aristocratie. »

Synonymes de *favoriser* sous son sens numéro 4 :
privilégier
avantager
soutenir

Dépendances impliquant *favoriser* :
`SUBJ(favorisé,règne)`
`VARG[DIR](favorisé,décadence)`

Dépendances générées par l'enrichissement :
`SUBJ(privilégier,règne)`
`SUBJ(avantager,règne)`
`SUBJ(soutenir,règne)`
`VARG[DIR](privilégier,décadence)`
`VARG[DIR](avantager,décadence)`
`VARG[DIR](favoriser,décadence)`

FIG. 5.8 – Enrichissement synonymique simple

L'exemple 5.8 illustre l'enrichissement synonymique du mot *favoriser* dans une phrase où son quatrième sens (*apporter de l'aide à une « notion abstraite »*) a été assigné par la désambiguïsation sémantique. Les synonymes correspondant à ce quatrième sens sont *privilégier*, *avantager* et *soutenir*. Deux dépendances seulement impliquent le lemme *favoriser* dans l'analyse syntaxique de l'énoncé. Étant donné qu'il y a trois synonymes de *favoriser*, ce sont six nouvelles dépendances qui sont créées, trois de chaque type dont l'argument correspondant à *favoriser* est remplacé successivement par chaque synonyme.

Cette méthode se révèle efficace dans divers tests que nous avons effectués, mais elle présente deux inconvénients majeurs : elle ne fonctionne réellement que si un seul des arguments des dépendances est enrichi, et elle génère un nombre de dépendances qui croît non seulement à mesure que la liste des synonymes est plus longue pour chaque unité lexicale, mais aussi en proportion du nombre de dépendances qui impliquent chaque unité lexicale enrichie.

Synonymes de *décadence* sous son sens numéro 1 :
déclin
corruption
déchéance

Dépendance impliquant *favoriser* et *décadence* :
`VARG[DIR](favorisé,décadence)`

Dépendances correspondantes générées par l'enrichissement de *décadence* :
`VARG[DIR](favorisé,déclin)`
`VARG[DIR](favorisé,corruption)`
`VARG[DIR](favorisé,déchéance)`

Dépendances manquantes après enrichissement :
`VARG[DIR](privilégier,déclin)`
`VARG[DIR](avantager,déclin)`
`VARG[DIR](soutenir,déclin)`
`VARG[DIR](privilégier,corruption)`
`VARG[DIR](avantager,corruption)`
`VARG[DIR](soutenir,corruption)`
`VARG[DIR](privilégier,déchéance)`
`VARG[DIR](avantager,déchéance)`
`VARG[DIR](soutenir,déchéance)`

FIG. 5.9 – Lacune de la méthode élémentaire d'enrichissement synonymique simple.

Nous présentons dans l'exemple 5.9 une illustration des critiques adressées à la méthode élémentaire d'enrichissement. Tout d'abord, pour une seule dépendance reliant *favoriser* et décadence, on obtient six dépendances différentes supplémentaires [10], soit une par synonyme d'un argument. Ensuite, aucune des dépendances ainsi générées ne permet de relier deux unités lexicales apportées par l'enrichissement synonymique, ce qui laisse dans le cas présent neuf dépendances possibles inexploitées. Le volume informationnel est donc important, mais l'information est toutefois insuffisante par rapport à ce qu'elle pourrait être.

Il s'agit dès lors d'ajouter une information synonymique à l'intérieur de chaque dépendance impliquant le mot à enrichir sans la dupliquer à chaque

[10] Trois de ces dépendances sont présentées dans l'exemple 5.8 page précédente.

fois, pour disposer non seulement de l'ensemble des synonymes d'un mot dans une même dépendance, mais aussi pour permettre toutes les compositions d'arguments au sein de la dépendance. Une disjonction entre le mot à enrichir et chacun de ses synonymes à l'intérieur de la dépendance permettrait de réaliser ces compositions à travers une seule dépendance. Toutefois, le formalisme de *XIP* se pose ici comme obstacle principal à cette réalisation. En effet, il n'autorise pas d'alternative dans l'expression des arguments d'une dépendance.

Cependant, il est possible d'emmagasiner une dépendance contenant des arguments disjonctifs comme s'il s'agissait d'une expression correspondant au formalisme *XIP*. Le système de stockage et d'indexation des réalités extraites de la base textuelle ou apportées par l'enrichissement permet en effet de présenter une structure à plat dans laquelle les différentes informations ne sont pas cataloguées les unes par rapport aux autres. Cette structure plate contient donc la dépendance, les mots qui en constituent les arguments et leurs enrichissements, mais rien ne détermine les rapports que ces différentes informations entretiennent les unes vis-à-vis des autres. En passant à un niveau inférieur – le niveau de la dépendance – il est possible de reconstituer l'ensemble des dépendances possibles seulement en cas de besoin lors de la phase de mise en correspondance de la question et des candidats réponse [11].

VARG[DIR]	privilégier	corruption
favoriser	avantager	déclin
soutenir	déchéance	décadence

FIG. 5.10 – Structure plate contenant les données correspondant à la dépendance enrichie.

Dans l'exemple 5.10, on peut voir que la structure plate permet de détecter la présence d'unités lexicales et de dépendances [12] dans un fragment de texte. Le fait de repérer les éléments recherchés dans la structure plate correspondant à une phrase déclenche une recherche au niveau de la dépendance. À ce niveau, c'est la dépendance disjonctive qui a été stockée (*cf.* figure 5.11 page suivante). Son aspect formel permet de reconstituer aisément dans un format conforme à l'analyse de *XIP* toutes les dépendances correspondant aux compositions d'arguments proposés par l'énoncé original

[11] Ce mécanisme sera explicité plus en détail au chapitre 6 page 183 qui est consacré à l'interrogation des documents.

[12] Par souci de clarté, nous n'avons indiqué que le nom de la dépendance et les unités lexicales. Dans la structure réelle, on trouve tous les traits qui y ont été associés et les autres informations recueillies lors de l'analyse de l'énoncé, ainsi que les indications permettant de reconstituer les dépendances.

et ses enrichissements.

$$\texttt{VARG[DIR]} \left(\begin{array}{c} \texttt{favoriser} \\ \text{OU} \\ \texttt{privilégier} \\ \text{OU} \\ \texttt{avantager} \\ \text{OU} \\ \texttt{soutenir} \end{array} , \begin{array}{c} \texttt{décadence} \\ \text{OU} \\ \texttt{corruption} \\ \text{OU} \\ \texttt{déclin} \\ \text{OU} \\ \texttt{déchéance} \end{array} \right)$$

FIG. 5.11 – Présentation disjonctive d'une dépendance enrichie.

Avec cette présentation de l'enrichissement par synonymie, nous gardons un accès total à l'information que nous ajoutons à l'analyse textuelle simple sans augmenter démesurément l'espace nécessaire à son stockage. Toutefois, une telle procédure ne peut fonctionner pour les enrichissements effectués au travers d'une expression synonymique constituée de plusieurs unités lexicales.

5.4.2 Enrichissement par expressions synonymiques

Le problème qui se pose lors d'un enrichissement avec expressions synonymiques vient du fait qu'elles sont composées de plusieurs unités lexicales et que de ce fait leur analyse syntaxique est complexe. Dans une tentative où la méthode d'enrichissement par ces expressions reste la même que celle qui exploite les synonymes simples, on obtient une ou plusieurs dépendances dont un des arguments au moins est une expression à mots multiples. Cependant, l'analyse syntaxique d'une phrase contenant la même expression dans son énoncé n'aboutit pas à une dépendance contenant la même expression, car celle-ci est analysée comme une suite d'unités lexicales et analysée comme telle. L'enrichissement d'un énoncé à travers une expression synonymique ne doit donc pas être réalisé selon la procédure utilisée pour les synonymes simples sous peine de n'être pas exploitable.

L'exemple 5.12 page ci-contre montre un énoncé où le lexème *commander* est désambiguïsé sous son troisième sens, ce qui permet un enrichissement par l'expression synonymique *exercer son autorité sur*. Toutefois, le remplacement élémentaire de *commander*, dans toutes les dépendances où il apparaît, par *exercer son autorité sur* ne permet pas d'obtenir des dépendances compatibles avec celles qui résultent d'une phrase où apparaîtrait la même expression[13]. Cette méthode est donc inopérante pour les expressions

[13]Aucune phrase de notre corpus ne présentant cette expression, nous avons analysé une phrase formelle *X exerce son autorité sur Y* où *X* et *Y* sont des noms. Le résultat de

Énoncé à enrichir :
« Il **commandait** les légions de Germanie. »

expression synonymique de *commander* au sens n°3, « gérer quelqu'un » :
exercer son autorité sur

Dépendances impliquant *commander* :
`SUBJ(commandait,Il)`
`VAR[DIR](commandait,légions)`

Dépendances issues d'un enrichissement élémentaire :
`SUBJ`(**exercer son autoritésur**, `Il`)
`VARG`[`DIR`](**exercer son autorité sur**, `légions`)

Analyse syntaxique de l'expression *exercer son autorité sur* dans un énoncé :
`SUBJ(exerce,X)`
`VARG[DIR](exerce,autorité)`
`VARG[INDIR](exerce,sur,Y)`

FIG. 5.12 – Problèmes liés à un enrichissement élémentaire par expression synonymique.

synonymiques de même que la méthode de construction de dépendances disjonctives.

Pour adapter l'information d'enrichissement apportée par une expression synonymique à mots multiples à un schéma compatible avec la structure dans laquelle nous emmagasinons toutes les réalités extraites de la base documentaire ou générées à partir d'elle, il s'agit de fournir une analyse linguistique de l'expression synonymique et des rapports que les unités lexicales qui la composent ont avec le reste de la phrase lorsqu'elle remplace dans le texte l'unité lexicale dont elle est synonyme. La méthode que nous appliquons consiste simplement à reconstruire la phrase concernée en remplaçant le mot à enrichir par son expression synonymique, puis à analyser la phrase ainsi générée selon les mêmes modalités que l'énoncé original.

Ainsi, après qu'une phrase présentée à notre système a été analysée par *NTM-XIP* et désambiguïsée, commence pour elle la phase d'enrichissement initiée par les processus liés à la synonymie. Les enrichissements liés aux synonymes simples sont effectués sur l'ensemble de la phrase en parallèle dans une étape précédant l'adjonction d'expressions synonymiques. Ensuite, chaque

l'analyse de cette phrase est présenté dans l'exemple 5.12.

expression synonymique disponible donne lieu à la génération d'une nouvelle phrase. Pour réaliser cette génération, l'unité lexicale originale est remplacée successivement par chacune de ses expressions synonymiques. Chaque nouvelle phrase est ensuite analysée, désambiguïsée et enrichie – excepté les unités lexicales appartenant à l'expression synonymique qui enrichit la phrase originale. Ensuite, les dépendances syntaxiques issues de la nouvelle analyse et redondante par rapport aux résultats de l'analyse et de l'enrichissement de la phrase originelle sont éliminées tandis que les autres sont conservées comme enrichissement de la structure syntaxico-sémantique de la base documentaire.

Il est fréquent que dans un même énoncé, plusieurs unités lexicales possèdent un lien de synonymie avec une expression complexe. Lorsque c'est le cas, autant de nouvelles phrases sont générées qu'il y a de combinaisons possibles des lexèmes originaux et des expressions synonymiques qui les enrichissent entre les unités lexicales à enrichir.

L'exemple présenté dans les figures 5.13 page ci-contre, 5.14 page 178 et 5.15 page 179 montre la complexité qu'il y a à enrichir un texte à l'aide d'expressions synonymiques. Dans un premier temps (figure 5.13 page suivante), l'enrichissement par synonymes simples est effectué selon la méthode que nous avons expliquée plus haut, et les dépendances qui en résultent sont stockées provisoirement dans une liste.

Ensuite (figure 5.14 page 178), toutes les unités lexicales dont l'enrichissement réclame l'exploitation d'une expression synonymique sont successivement remplacées par cette expression synonymique pour former à chaque fois un nouvel énoncé. Cet énoncé est à son tour analysé et enrichi de la même manière que la phrase originale, excepté les unités lexicales qui font partie de l'expression synonymique. Les dépendances obtenues à la fin de l'enrichissement sont comparées à celles de la liste provisoire des dépendances obtenues pour la phrase originale ainsi que pour les énoncés construits à partir de cette phrase. Les dépendances originales[14] sont conservées tandis que les autres sont éliminées.

Enfin, toutes les combinaisons d'enrichissements par expressions synonymiques sont tentées pour construire une nouvelle phrase à partir de l'énoncé original (figure 5.15 page 179). Pour chaque combinaison différente, un nouvel énoncé est construit, analysé et enrichi, et les nouvelles dépendances sont emmagasinées tandis que les doublons sont rejetés. Lorsque toutes les combinaisons ont été testées, la liste provisoire contient l'ensemble de l'information d'enrichissement synonymique de l'énoncé. Elle est alors intégrée à la structure syntaxico-sémantique de la base textuelle tandis que le processus passe à un type d'enrichissement ultérieur.

[14] L'exemple signale que ces dépendances sont originales en les écrivant en caractères gras.

Énoncé à enrichir :
« Il **commandait** les **légions** de Germanie. »

Synonymes pour *commander* (sens n°3, « gérer quelqu'un ») et *légion* (sens n°1, « armée romaine ») :

contrôler	troupe
exercer son autorité sur	armée
	unité militaire

Dépendances syntaxiques extraites par *NTM-XIP* :

```
SUBJ(commande,il)
VARG[DIR](commande,légions)
NMOD[INDIR](légions,de,Germanie)
```

Enrichissement par synonymes simples :

$$\mathbf{SUBJ}\left(\begin{array}{c}\mathbf{commande}\\ \mathbf{OU}\\ \mathbf{contrôler}\end{array}\ ,\ \mathbf{Il}\right)$$

$$\mathbf{VARG[DIR]}\left(\begin{array}{c}\mathbf{commande}\\ \mathbf{OU}\\ \mathbf{contrôler}\end{array}\ ,\ \begin{array}{c}\mathbf{troupe}\\ \mathbf{OU}\\ \mathbf{armée}\\ \mathbf{OU}\\ \mathbf{légion}\end{array}\right)$$

$$\mathbf{NMOD[INDIR]}\left(\begin{array}{c}\mathbf{troupe}\\ \mathbf{OU}\\ \mathbf{armée}\\ \mathbf{OU}\\ \mathbf{légion}\end{array}\ ,\ \mathbf{de}\ ,\ \mathbf{Germanie}\right)$$

FIG. 5.13 – Enrichissement simple d'un énoncé présentant des possibilités d'enrichissement par expressions synonymiques.

Cependant, nous n'avons pas abordé le problème de l'analyse textuelle des nouveaux énoncés. En effet, ils sont créés non pas avec une forme fléchie mais avec la **forme canonique** de l'expression synonymique proposée pour l'enrichissement telle qu'elle apparaît dans le dictionnaire. Pour que l'analyse morpho-syntaxique des nouveaux énoncés ne pose pas de problème, nous soumettons chaque expression synonymique à une analyse morphologique afin de déterminer dans l'expression la première unité lexicale de même catégorie grammaticale que le lexème à enrichir. Nous pouvons dès lors transmettre les traits morphologiques du lexème à enrichir à ce que nous considérons comme la tête de l'expression synonymique, ce qui permet de contourner les problèmes d'analyse syntaxique liés à une analyse morphologique erronée ou

• Enrichissement de *commander* par expression synonymique :
Nouvel énoncé :
« Il **exercer son autorité sur** les légions de Germanie. »

Dépendances enrichies :
SUBJ(**exerce**, **Il**)
VARG[**DIR**](**exerce**, **autorité**)

$$\textbf{NMOD}[\textbf{INDIR}]\left(\begin{array}{ccccc} & & & & \textbf{troupe} \\ & & & & \textbf{OU} \\ \textbf{exercer} & , & \textbf{sur} & , & \textbf{armée} \\ & & & & \textbf{OU} \\ & & & & \textbf{légion} \end{array}\right)$$

$$\texttt{NMOD[INDIR]}\left(\begin{array}{ccccc} \texttt{troupe} & & & & \\ \text{OU} & & & & \\ \texttt{armée} & , & \texttt{de} & , & \texttt{Germanie} \\ \text{OU} & & & & \\ \texttt{légion} & & & & \end{array}\right)$$

• Enrichissement de *légion* par expression synonymique :
Nouvel énoncé :
« Il commande les **unité militaire** de Germanie. »

Dépendances enrichies :

$$\texttt{SUBJ}\left(\begin{array}{ccc} \texttt{commande} & & \\ \text{OU} & , & \texttt{Il} \\ \texttt{contrôler} & & \end{array}\right)$$

$$\textbf{VARG}[\textbf{DIR}]\left(\begin{array}{ccc} \textbf{commande} & & \\ \textbf{OU} & , & \textbf{unité} \\ \textbf{contrôler} & & \end{array}\right)$$

NMOD[**ADJ**](**unité**, **militaire**)
NMOD[**INDIR**](**unité**, **de**, **Germanie**)

FIG. 5.14 – Enrichissement par une expression synonymique dans un nouvel énoncé.

à une mauvaise désambiguïsation catégorielle.

De cette manière, dans l'exemple précédent (figure 5.15 page ci-contre), *exercer* est considéré comme un verbe à l'indicatif présent, troisième personne du singulier dès avant son analyse morphologique, ce qui permet de lui donner une dépendance de type sujet qui n'apparaîtrait pas si le même énoncé était analysé avec la forme infinitive décelée par l'analyse morphologique. Après avoir résolu les différents problèmes rencontrés au cours des tests, nous sommes maintenant en mesure d'effectuer correctement un en-

•Enrichissement de *commander* et *légion* par expression synonymique :
Nouvel énoncé : « Il **exercer son autorité sur** les **unité militaire** de Germanie. »
Dépendances enrichies :
`SUBJ(exercer,il)`
`VARG[DIR](exercer,autorité)`
VARG[INDIR](exercer, sur, unité)
`NMOD[ADJ](unité,militaire)`
`NMOD[INDIR](unité,de,Germanie)`

FIG. 5.15 – Combinaison des enrichissements par expressions synonymiques dans un seul énoncé.

richissement à l'aide des synonymes. D'autres enrichissements peuvent dès lors être envisagés.

5.5 Exploitation de la dérivation morphologique

Lors de notre examen des ressources lexico-sémantiques, nous avons signalé diverses informations qui nous paraissaient aptes à apporter un enrichissement à des énoncés textuels. Parmi des données d'enrichissement textuel, nous avons déjà traité les divers traits sémantiques (domaines, classes et catégories) en élargissant le contenu du lexique morphologique, ainsi que l'information synonymique dont nous venons de décrire la méthodologie d'adjonction à la structure syntaxico-sémantique de la base documentaire.

Nous avons par ailleurs fait mention des indications de dérivation morphologique contenues dans le dictionnaire *Dubois*. De plus, nous avons été confronté au problème de la génération des dérivés, que nous avons résolue grâce à un outil de morphologie relationnelle déjà existant [Gaussier, 1999] (*cf.* section 3.3 page 116). Enfin, nous avons étudié l'évolution du sens des différentes dérivations pour en déduire des tables de correspondances syntaxiques permettant de faire coïncider avec une forme originelle dans un texte une forme dérivée dans ce même texte (*cf.* section 4.2.2 page 133). Il nous reste à étudier la manière dont nous allons exploiter ces dérivés que nous avons générés et ces tables de correspondance pour enrichir la structure syntaxico-sémantique.

L'exemple 5.16 page suivante illustre le fonctionnement d'un schéma syntaxique extrait des tables de correspondances. Ainsi, dans la phrase *Pline le Jeune protégea Suétone*, le mot *protecteur* peut être proposé comme déri-

Énoncé présenté à l'analyse syntaxique :
Pline le Jeune **protégea** Suétone.

Forme dérivée :
protecteur

Correspondance syntaxique :
```
VARG[DIR](verbe,X)
     ==> NMOD[INDIR](nom dérivé, PREP, X)
```

Évolution syntaxique de l'énoncé :
```
VARG[DIR](protéger, Suétone)
     ==> NMOD[INDIR](protecteur, PREP, Suétone)
```

FIG. 5.16 – Application d'une correspondance syntaxique pour un dérivé de *protéger*.

vation du verbe *protéger*. L'étude des dépendances obtenues lors de l'analyse de cet énoncé par *XIP* permet d'identifier dans les tables la correspondance entre une relation syntaxique de type objet direct entre un verbe et une autre unité lexicale, et une relation prépositionnelle entre le dérivé de ce verbe et l'autre unité lexicale. Cette correspondance syntaxique permet d'établir une conformité sémantique entre *protéger Suétone* et *protecteur [de] Suétone* [15]. Toutefois, une telle conformité syntaxique ne s'exprime pas au niveau d'un énoncé, mais seulement au niveau de la dépendance `NMOD[INDIR](protecteur,PREP,Suétone)`.

Ici encore, l'information indicative des possibilités de dérivation morphologique est distribuée non en fonction du mot-vedette, mais suivant les acceptions de cette vedette. Toute la procédure qui vise à un enrichissement par les dérivés est donc soumise à nouvaus au bon déroulement de la désambiguïsation sémantique. Par ailleurs, on a vu que l'intégration d'une forme dérivée à la structure syntaxico-sémantique ne peut se faire la plupart du temps que *via* une transformation du contexte syntaxique de la forme originale lorsque intervient le dérivé. Dès lors, l'application d'un enrichissement par dérivation à une unité lexicale est soumise à deux conditions : la sélection d'un sens de ce lexème qui préconise une dérivation, et la détection d'un des schémas syntaxiques permettant une adaptation correspondante de l'énoncé.

[15] Ou bien *protecteur pour Suétone*, ou encore *protecteur envers Suétone*, le lemme de la préposition n'étant pas spécifié.

Une fois vérifiée la validité de la génération d'une forme dérivée pour le sens sélectionné de l'unité lexicale à enrichir, ainsi que la conformité du contexte syntaxique original de cette unité lexicale avec la table des correspondances syntaxiques propre à ce type de dérivé, la procédure d'enrichissement par forme dérivée consiste à construire le schéma syntaxique correspondant à partir de ce contexte syntaxique original. Les dépendances syntaxiques correspondant au schéma initial et celles qui en sont issues suivant les indications de la table des correspondances constituent l'ossature de deux expressions de même sens.

Le schéma syntaxique dérivé représente dès lors un enrichissement plus ou moins paraphrastique apporté à l'énoncé de départ, qui sera versé dans la structure syntaxico-sémantique au même titre que les résultats de l'analyse du même énoncé original ou que les informations provenant des autres enrichissements. Un trait identifiant le type de cet enrichissement est toutefois assigné aux dépendances créées suivant les directives de la table des correspondances syntaxiques.

Le système de base de données dans lequel sont stockées les informations issues de l'analyse des textes et de l'enrichissement occupe un espace bien plus important que celui des textes originaux. Le tableau 5.5 donne les détails de l'espace occupé par les données issues de 50 articles de l'*Encyclopédie Hachette Multimédia* utilisés au chapitre 7 page 197 pour effectuer l'évaluation du système. Ces textes occupent 0,2 Mo.

	Analyse syntaxique		Tous enrichissements	
	Volume	# lignes	Volume	# lignes
Données langagières	2,8 Mo	9 132 l.	4 Mo	13 719 l.
Structures hiérarchiques	6 Mo	172 218 l.	10 Mo	324 314 l.
Structures plates	2,8 Mo	54 382 l.	4,5 Mo	90 720 l.
Niveau phrase	1,6 Mo	31 546 l.	2,49 Mo	50 079 l.
Niveau paragraphe	0,9 Mo	16 929 l.	1,45 Mo	29 272 l.
Niveau document	0,3 Mo	5 907 l.	0,56 Mo	11 369 l.
Total	11,6 Mo	235 732 l.	18.5 Mo	428 753 l.

TAB. 5.2 – Espace relatif occupé par la structure informationnelle d'une base documentaire.

La ligne *données langagières* du tableau représente l'index de l'ensemble des données extraites des documents, c'est-à-dire les unités lexicales, les dépendances et les traits. Les structures hiérarchiques sont les index de chaque dépendance extraite avec ses traits, ses arguments et les traits des arguments tels qu'ils apparaissent lors de l'analyse du texte ou lors d'une phase d'enrichissement. Les structures plates représentent ces informations qui ap-

paraissent au niveau de la phrase, du paragraphe ou du document après analyse ou après enrichissement, mais sans que les données soient structurées les unes vis-à-vis des autres. Les structures plates sont de simples listes des données contenues dans une phrase, dans un paragraphe ou dans un document (*cf.* section 6.2.2 page 187 et l'annexe A page 275). Le volume des données est indiqué après une simple analyse syntaxique et après tous les traitements d'analyse et d'enrichissement.

La grande importance de l'espace occupé par la structure informationnelle par rapport à la base documentaire provient de trois facteurs :

- un grand nombre de données sont ajoutées tant au cours de l'analyse que pendant l'enrichissement. Il est d'ailleurs possible que beaucoup de ces données ne soient pas utiles dans une structure de l'information (par exemple certains traits morphologiques peuvent être supprimés) ;
- l'architecture du système de base de données répond plus à des besoins immédiats de test qu'à des impératifs d'optimisation. Nous ne sommes pas en mesure de juger les possibilités de compactage des index ;
- les besoins de distinguer chaque type d'information et la provenance de chaque enrichissement nous a imposé de distinguer souvent des données identiques simplement parce que leur provenance était différente. Les tests effectués par Claude Roux sur des informations non distinguées montrent que, selon les corpus, l'espace utilisé pouvait se réduire de plus de dix fois.

5.6 Conclusion

Après avoir effectué une analyse morpho-syntaxique du texte qui s'est prolongée par une identification sémantique des unités lexicales, puis avoir procédé à un enrichissement lexical, syntaxico-sémantique et sémantique, nous avons emmagasiné une information considérable dans une structure qui constitue l'architecture informative de la base textuelle traitée. Cette structure syntaxico-sémantique présente la particularité de contenir une même information sous plusieurs formulations différentes en fonction de l'enrichissement qui a pu en être fait. En outre, l'assignation, sous forme de traits sémantiques, de caractéristiques plus abstraites telles que domaines, classes et catégories sémantiques permet dans une certaine mesure un détachement de la forme de surface.

Cette structure syntaxico-sémantique très riche nous met donc à même de manipuler l'information textuelle au niveau de son sens. Dès lors, nous pouvons aborder la phase de manipulation de cette information. Une des tâches les plus exigeantes dans ce domaine est celle de question-réponse. C'est au travers d'une application de type question-réponse que nous allons tester la qualité de la structure informationnelle que nous avons créée.

Chapitre 6

Interrogation de la base documentaire

6.1 Introduction

L'élaboration de la structure informationnelle d'une base documentaire n'est pas une réalisation gratuite. Sa vocation est de fournir un accès précis à l'information contenue dans les textes, même pour des disciplines aussi exigeantes que l'extraction d'information ou la tâche de question-réponse. Une fois la structure informationnelle construite, il s'agit de concevoir une méthode qui permette de sélectionner l'information jugée intéressante.

Comme la nature de l'information contenue dans la structure informationnelle est linguistique (lexèmes, dépendances syntaxiques, traits sémantiques, etc.), il est normal de demander à la méthode d'interrogation de la base d'être capable de traiter des requêtes en langage naturel. Dès lors, l'accès à l'information se fait en questionnant directement le système. Cette requête est analysée d'une manière similaire à l'analyse des documents afin de construire une structure correspondant à la question compatible avec la structure informationnelle de la base documentaire. Ensuite, la structure de la question est comparée avec celle des textes afin de localiser la présence d'une information correspondante dans les textes.

L'analyse de la question est particulière dans la mesure où il s'agit de catégoriser l'information recherchée. Par ailleurs, aucun enrichissement de la structure de la question n'est effectué. La comparaison entre les structures de la question et des documents peut être plus ou moins exigeante. D'abord, l'unité lexicale qui permet de catégoriser l'information recherchée peut être exigée ou non dans la réponse. Ensuite, la correspondance des liens syntaxiques entre les éléments d'information de la question et de la réponse peut être plus ou moins grande. Enfin, la fenêtre de réponse peut varier en

taille : phrase, paragraphe ou texte.

Ce chapitre décrit les traitements particuliers appliqués à la question pour construire sa structure informationnelle propre, et notamment la catégorisation de l'objet de la question. Ensuite, il présente la méthode de mise en correspondance de la question avec les fragments de texte qui en sont les réponses, ainsi que les paramètres de contrainte sur l'objet de la question, sur les correspondances syntaxiques et sur la fenêtre de réponse.

6.2 Analyse de la question

L'apport d'une réponse à une question ne peut se faire qu'en mettant en correspondance une information partielle présente dans l'énoncé de la question avec une information complète concordante dans la base documentaire que l'on interroge. Il s'agit dès lors d'extraire cette information partielle de la question sous une forme qui correspond à celle sous laquelle nous avons conservé les données de base et les enrichissements de la base documentaire à interroger. C'est donc sur la base d'une analyse semblable que nous devons traiter les questions posées au système.

Cette analyse ne sera cependant pas la même que celle qui est à l'origine de la construction de la structure de la base documentaire. En effet, les enrichissements ayant déjà été effectués lors de l'analyse des documents, il n'est en principe pas nécessaire de repasser par une telle procédure. De plus, le contexte de la question étant souvent bien plus restreint que celui d'un document réel, il est probable que les différentes méthodologies mises en œuvre au cours de notre approche ne pourraient prétendre à un même niveau d'exactitude.

Or chacun des enrichissements que notre système d'analyse a permis d'effectuer sur la base documentaire est potentiellement générateur de bruit dès lors que le résultat d'une analyse, tronqué par la limite de l'information contextuelle, présente des erreurs dans son interprétation. On se souviendra que c'est cette réserve quant à la richesse de l'information contextuelle de la requête qui nous a amené à considérer l'importance de l'enrichissement du document [1]. Toutefois, l'analyse de l'énoncé de la requête requiert certaines des adaptations que nous avons apportées au lexique fourni à *NTM* ainsi qu'à la grammaire de *XIP* pour l'analyse du document, ainsi que d'autres particularités qui tiennent à la nature interrogative des requêtes présentées au système.

[1] Malgré cela, il ne nous paraît pas dénué d'intérêt de tenter l'interrogation de la base documentaire enrichie au moyen de requêtes enrichies elles aussi. Nous comptions faire cette expérience au cours de l'évaluation du système au chapitre suivant et étudier l'impact que cela peut avoir sur les résultats, mais le temps nous a manqué pour mettre au point cette expérience.

6.2.1 Traitements communs aux documents et à la question

Les traitements qui sont communs au traitement des documents de la base et à l'énoncé des questions proposées au système sont bien entendu l'analyse morpho-syntaxique de *NTM-XIP*, qui permet d'obtenir une réelle cohérence entre les traits morphologiques et syntaxiques extraits de la question et des fragments de textes, ainsi qu'entre les dépendances générées. Comme pour l'analyse des documents, certains de ces traits et de ces dépendances ne sont pas pris en compte : la catégorie grammaticale mise à part, les indications morphologiques ne sont pas prises en considération ; les relations que nous avons appelées « fonctionnelles » (*cf.* section 5.3.1 page 157) sont également éliminées[2].

Certains des traitements appliqués aux questions posées au système sont donc les mêmes que ceux qui ont été utilisés pour les documents. Dès lors, nous ne nous attarderons pas à en décrire le fonctionnement, mais nous justifierons une nouvelle utilisation de ces méthodes.

Tout d'abord, nous soumettons les questions à l'analyse morpho-syntaxique de *NTM-XIP* afin de disposer d'informations morphologiques et syntaxiques cohérentes avec celles qui ont été indexées dans la structure syntaxico-sémantique de la base documentaire. De même qu'au cours de l'analyse des documents, nous ne conservons de l'information morphologique que des données relatives à la catégorie grammaticale des lexèmes.

Pareillement, les unités lexicales ne seront conservées que dans la mesure où elles apparaissent comme arguments de dépendances syntaxiques calculées par *XIP*. Et bien entendu, il est naturel de ne pas conserver les dépendances syntaxiques fonctionnelles qui ne peuvent correspondre à aucune des réalités emmagasinées dans la structure, puisqu'elles ont été rejetées lors de l'analyse des documents.

De plus, les lexiques utilisés pour effectuer cette analyse sont identiques à ceux qui ont été exploités lors de la construction de la structure documentaire. De ce fait, les traits sémantiques reprenant les domaines d'utilisation, les classes sémantiques et les catégories sémantiques, dont nous avons enrichi le lexique morphologique (*cf.* section 4.2.2 page 133), seront assignées aux éléments extraits de l'énoncé de la requête lors de son analyse. Cette assignation de traits permettra, dans une certaine mesure, d'effectuer un typage de l'objet de la question.

C'est à ce stade de l'analyse de la requête que le traitement commence à diverger de celui qui a été appliqué aux textes de la base documentaire. En ef-

[2]Même s'il est possible de conserver ces relations fonctionnelles tant dans la structure informationnelle que dans l'analyse de la question, en débrayant la fonction d'élimination.

fet, les autres traitements concernent purement l'enrichissement de l'énoncé, que nous avons renoncé à appliquer à un texte aussi court qu'une question. Il est toutefois important d'indiquer que dès ce niveau d'analyse, on dispose du squelette de la phrase sur lequel est construit toute la méthodologie d'enrichissement et de construction de la structure syntaxico-sémantique de la base textuelle.

6.2.2 Divergences dans la méthode d'analyse

La première des libertés que prend le traitement de la requête par rapport à celui des documents réside dans un apport que nous avons fait à la grammaire et qui permet d'identifier l'objet de la question dans une certaine mesure – ou en tout cas de le catégoriser. En effet, pour pouvoir apporter une réponse à la question posée au système, il importe d'identifier le plus précisément possible les caractéristiques attachées à l'élément capable d'y répondre.

Apport lexical

Or comme les unités lexicales – pronoms, adjectifs ou adverbes – permettant d'introduire une interrogation n'existent qu'en nombre limité, et que leur nature permet de catégoriser souvent la réponse qu'ils attendent, nous avons ajouté à la grammaire du français de *XIP* plusieurs règles lexicales qui permettent d'attacher des traits sémantiques aux interrogatifs et ainsi de fixer certaines contraintes sur les réponses candidates fournies par le système.

Nous aurons par exemple un trait humain qui s'attachera au pronom interrogatif *qui*, des traits de `temps` ou de `lieu` qui marqueront respectivement les adverbes interrogatifs *quand* et *où* [3]. Par contre, certaines unités lexicales interrogatives ne peuvent être catégorisées de cette manière. Ainsi, *que* ou *quoi* ne peuvent recevoir systématiquement de trait sémantique [4], et les traits portant sur les adjectifs interrogatifs sont susceptibles de varier en fonction du lexème qu'ils qualifient.

Nous avons donc réalisé autant que possible des règles lexicales *XIP* attribuant certains traits sémantiques aux interrogatifs suffisamment typés pour

[3] On se souviendra que cette application ne permet pas de demander au système de porter des jugements. Dès lors, les interrogations introduites par *pourquoi* et *comment* ne sont pas permises. En effet, de nombreuses questions introduites par ces interrogatifs appellent un jugement. La création d'un module permettant de déterminer si la question appelle un jugement ou non requerrait un temps dont nous ne disposions pas, et de plus ajoutait une procédure à tester au cours de l'évaluation qui n'appartenait pas à notre recherche.

[4] Tout au plus peut-on inhiber pour ces interrogatifs le trait *humain*.

recevoir une telle affectation. Il a toutefois fallu trouver un autre procédé pour permettre la catégorisation de l'objet des questions dont l'interrogatif ne permet pas une telle détermination.

Une identification syntaxique : le FOCUS

Le ***focus*** d'une question est une notion introduite par [Lehnert, 1979]. Il correspond dans cet ouvrage à un concept présent dans la question qui englobe l'information attendue en réponse à cette question. Largement reprise et redéfinie par la suite, elle est pour [Ferret et al., 2002b] un mot ou un groupe nominal de la question qui représente le concept sur lequel une information est demandée par la question, et qui se trouve habituellement dans la réponse. De notre point de vue, le *focus* correspond à l'objet de la question. Il s'agit d'une unité lexicale qui détermine à l'intérieur de la question les caractéristiques sémantiques de ce que doit être la réponse. De plus, sa fonction syntaxique dans la question n'est pas quelconque : le *focus* entretient avec l'interrogatif une relation privilégiée quand il n'est pas lui-même l'interrogatif[5].

Nous avons donc créé un nouveau type de dépendance qui ne correspond à aucune relation syntaxique traditionnelle, mais prend en argument l'objet de la question. Ainsi, si c'est un adjectif qui introduit la requête et est porteur de sa fonction interrogative, l'unité lexicale dont il est épithète sera l'argument de cette dépendance, que nous avons appelée FOCUS. Par contre, lorsque c'est un pronom qui introduit la question, deux possibilités se présentent : si ce pronom est le sujet d'un verbe copule, le FOCUS portera sur son attribut ; dans le cas contraire, c'est le pronom interrogatif lui-même qui sera l'argument du FOCUS (*cf.* figure 6.1 page suivante).

Bien qu'elle n'apparaisse pas dans la structure de la base documentaire, cette dépendance est pourtant très importante pour le bon fonctionnement de la procédure d'interrogation. En effet, quoiqu'un certain typage de la réponse attendue soit possible d'un point de vue lexical grâce aux lexèmes interrogatifs qui n'existent qu'en nombre limité (*cf. supra*), la plupart des interrogations ne sont réellement catégorisées qu'à la faveur du contenu de la question, et notamment le contexte syntaxique immédiat de l'interrogatif qui introduit cette question. La dépendance FOCUS a pour mission d'identifier l'unité lexicale qui constitue dans la question la plus grande détermination relative à la réponse.

[5]Nous distinguons dans la typographie le FOCUS, qui est la dépendance extraite pas la grammaire de XIP, et *focus*, qui est l'objet de la question.

FOCUS sur le nom dont l'interrogatif est épithète :
Quelle **ville** devint capitale de l'Empire Romain en 402 ?
`FOCUS(ville)`

FOCUS sur l'attribut de l'interrogatif :
Qui était le **beau-père** de Galère ?
`FOCUS(beau-père)`

FOCUS sur l'interrogatif :
Contre **qui** Constant Ier lutta-t-il ?
`FOCUS(qui)`

FIG. 6.1 – Exemples des différents types de dépendance `FOCUS`.

6.2.3 Exploitation des particularités de l'analyse des questions

Malgré les travaux effectués sur le typage de l'objet de la question, que ce soit au niveau lexical par détermination du type demandé par l'interrogatif introducteur de la requête ou du point de vue syntaxique avec l'unité lexicale constituant un contexte déterminant de l'interrogatif, encore faut-il être à même d'exploiter les informations que l'ensemble de cette analyse fournit. Or pour faire correspondre l'information que nous avons ainsi extraite de la question avec celle qui est contenue dans la structure syntaxico-sémantique de la base documentaire, la forme de cette information doit être compatible.

Le premier motif d'incompatibilité apparaît dans la présence d'un interrogatif dans les arguments des dépendances qui constituent le squelette de l'information extraite de la requête. Il est en effet bien rare que, dans une base textuelle élaborée pour contenir de l'information comme l'est l'ensemble des textes constitutifs d'une encyclopédie, on trouve des énoncés interrogatifs. Les unités interrogatives sont donc peu fréquentes et cette carence rend la plupart des dépendances des questions où intervient un interrogatif caduques pour leur mise en correspondance avec la base textuelle.

Pour éliminer cette incohérence entre les deux structures informationnelles, nous avons systématiquement supprimé l'unité lexicale interrogative, tout en maintenant le cas échéant les contraintes sémantiques qui lui étaient liées depuis l'application des règles lexicales *XIP*. Ainsi, une dépendance semblable mais présentant pour ce même argument n'importe quel lemme sera mise en correspondance avec la dépendance extraite de la question, à condition que ce lemme possède les mêmes traits sémantiques que l'interrogatif.

Qui persécutait les chrétiens ? (1)
(...) le monarque Châhpuhr II qui persécutait les chrétiens (...) (2)

Structure de (1)	Structure de (2)
SUBJ(**persécutait**, `Qui`[***humain*** *:+*])	**SUBJ**[`REL`](**persécutait**, `monarque`[***humain*** *:+*])
VARG[**DIR**](**persécutait**, **chrétiens**)	**VARG**[**DIR**](**persécutait**, **chrétiens**)
`FOCUS(Qui)`	`NN[PROPER](Châhpuhr,II)`
	`NN(monarque,Châhpuhr)`

FIG. 6.2 – Mise en correspondance d'une ***question*** avec une réponse candidate.

On peut voir dans l'exemple 6.2 que la structure informationnelle de la question ne peut pas s'apparier directement à celle de la réponse. La suppression de l'interrogatif est nécessaire car il n'est pas présent dans la réponse. Si l'identification du *focus* permet de catégoriser l'objet de la question et de remplacer l'unité lexicale qui est l'argument de cette dépendance (*Qui*) par sa catégorie sémantique (*humain*), il faut ensuite éliminer la dépendance `FOCUS`. Nous avons indiqué en caractères gras dans l'exemple les éléments de la structure informationnelle de la question qui sont maintenus après le traitement particulier de la question et permettent d'apparier la réponse avec la question.

Malgré cette suppression des interrogatifs, l'information extraite des requêtes n'est pas pleinement exploitable en l'état. En effet, le `FOCUS` est une dépendance qui n'apparaît qu'au cours de l'analyse des requêtes. Or, de ce fait, cette dépendance ne peut contribuer à une mise en correspondance de l'information de la question et celle des textes de la base documentaire. Et pourtant la catégorisation de l'objet de la réponse que cette dépendance effectue est capitale pour pouvoir apporter une réponse.

De la même manière que les règles lexicales liées aux interrogatifs apportent une catégorisation de la réponse attendue par les requêtes, la dépendance `FOCUS` est en mesure de fournir une information sémantique liée à l'objet de la réponse. Lorsque l'argument du *focus* est l'interrogatif lui-même, la dépendance fait double emploi, puisque ce sont les traits fournis par les règles lexicales qui sont appliquées pour assigner des traits sémantiques à l'interrogatif. Cette dépendance `FOCUS` est donc rejetée sans autre traitement.

Par contre, lorsqu'il s'agit d'une autre unité lexicale, ce sont les traits de ce mot qui sont maintenus comme condition à l'application de la règle. Toutefois, le lemme correspondant à ce lexème et son numéro de sens sont eux-mêmes des informations qui catégorisent la réponse attendue par la question. Ils sont donc maintenus eux aussi, mais nous y adjoignons un trait

`objetQuestion`[6], qui apparaîtra dans toutes les dépendances impliquant cette unité lexicale comme argument. Ce trait a pour fonction de signaler que l'unité lexicale qui la porte a été considéré comme le *focus* de la question. Nous verrons dans la section consacrée à la mise en correspondance de la question et des réponses candidates de quelle manière nous pourrons l'utiliser. Quant à la dépendance `FOCUS`, elle sera simplement éliminée une fois ces traitements effectués sur les autres dépendances.

Nous avons à présent extrait une importante information de la question posée au système, et nous l'avons manipulée de manière à ce qu'elle puisse être mise en correspondance avec celle qui est contenue dans la structure syntaxico-sémantique que nous voulons interroger. Il nous reste à décrire de quelle manière nous allons opérer pour apparier les réponses candidates et les questions.

6.3 Mise en correspondance des réponses avec la question

C'est dans un but d'uniformisation que nous avons appliqué un même type d'analyse aux documents et à la requête. En effet, cette uniformité a pour but de comparer les éléments informationnels extraits de cette requête avec ceux emmagasinés dans la structure syntaxico-sémantique de la base documentaire. L'ensemble de cette information est présentée sous la forme de dépendances, d'arguments de dépendances et de traits portant soit sur les dépendances, soit sur leurs arguments. La méthode de mise en correspondance d'une réponse avec sa requête consiste à retrouver dans la structure de la base documentaire un texte ou un fragment de texte qui contient les éléments informationnels de cette requête ainsi que l'élément recherché tel qu'il a été catégorisé. La structure de la question devra donc se retrouver entièrement dans celle du fragment de texte qui en constitue la réponse. De plus, cette information devra être agencée de la même manière dans le texte et dans la question.

6.3.1 Des structures plates parfaitement compatibles

La méthode d'indexation et de stockage de l'information que nous utilisons [Roux et Jacquemin, 2002] permet deux présentations différentes de l'information à chaque niveau de segmentation du document (texte, paragraphe, phrase, dépendance) : soit il s'agit de la structure hiérarchique permettant d'identifier chaque élément d'information par rapport aux autres,

[6] Ce trait ne possède pas de valeur intrinsèque, il est uniquement utilisé de manière fonctionnelle, pratique, pour signaler que l'unité lexicale qui le porte correspond au *focus*.

soit il s'agit de la structure « à plat » qui énumère les éléments informatifs sans expliciter les liens qu'ils ont les uns avec les autres (*cf.* annexe A page 275). Le parcours et la comparaison des structures plates est extrêmement rapide même s'il est peu approfondi. La comparaison des structures plates de la base documentaire avec celle de la requête va nous permettre d'éliminer le texte qui ne contient pas l'information contenue dans cette requête.

Qui persécutait les chrétiens ? (1)
Le monarque (...) persécutait les chrétiens (...) (2)
Les chrétiens persécutent le monarque (3)
Théodose défendit les chrétiens (...) (4)

Structure de (1)
`SUBJ(persécutait,?[humain :+])`
`VARG[DIR](persécutait,chrétiens)`

Structure de (2)
`SUBJ(persécutait,`
`monarque[humain :+])`
`VARG[DIR](persécutait,chrétiens)`

Structure de (3)
`SUBJ(persécutent,`
`chrétiens[humain :+])`
`VARG[DIR](persécutent,monarque)`

Structure de (4)
`SUBJ(défendit,`
`Théodose[humain :+])`
`VARG[DIR](défendit,chrétiens)`

Structures plates des énoncés

(1)	(2)	(3)	(4)
SUBJ	SUBJ	SUBJ	SUBJ
VARG	VARG	VARG	VARG
DIR	DIR	DIR	DIR
persécuter	persécuter	persécuter	**défendre**
chrétien	chrétien	chrétien	chrétien
	monarque	monarque	Théodose
humain	humain	humain	humain

FIG. 6.3 – Utilisation de la structure plate pour un filtrage des réponses candidates.

L'exemple 6.3 illustre le principe de création et d'utilisation de la structure plate d'un énoncé. Cette structure plate correspond à la liste des éléments d'information de la structure informationnelle de cet énoncé. Dans l'exemple, on peut voir l'intérêt de cette structure plate, qui permet, par une comparaison simple et rapide, l'élimination d'une réponse inexacte (énoncé (4) qui ne contient pas *persécuter*). On peut également en voir les limites avec l'énoncé (3), qui contient rigoureusement la même information que l'énoncé (2), mais qui agence cette information différemment. L'exploitation de la structure plate permet dès lors d'effectuer un tri et d'éliminer les phrases qui ne contiennent pas l'information requise. Il faut toutefois en passer par la struc-

ture informationnelle complète pour décider si les candidates qui passent ce premier filtre correspondent bien à la question posée.

Lorsque les réponses candidates ont été isolées, il faut vérifier que les liens entre les éléments informationnels correspondent à la syntaxe de la requête. Cette comparaison porte sur la nature des dépendances et sur les arguments des dépendances, les traits étant laissés de côté à ce stade de l'opération [7]. La vérification la plus aisée et la moins coûteuse porte sur les dépendances simples, issues de l'analyse syntaxique, de l'enrichissement synonymique simple ou de l'enrichissement par dérivations morphologiques. En effet, ces dépendances sont directement disponibles dans la structure informationnelle hiérarchique maintenant exploitée. En cas de succès de cette comparaison, la réponse candidate est considérée comme une réponse pertinente à la question et présentée à l'utilisateur. Le processus passe alors au traitement d'une éventuelle autre réponse candidate.

À défaut d'une correspondance complète des deux structures informationnelles [8], la procédure d'interrogation préconise de compléter la structure de la réponse candidate avec les dépendances disjonctives. Cette phase demande un traitement particulier car les dépendances disjonctives ne peuvent être directement comparées à celles qu'extrait l'analyseur *NTM-XIP* qui a généré la structure informationnelle de la question.

À partir de chaque dépendance disjonctive, il s'agit de reconstituer toutes les dépendances de même nature que la dépendance disjonctive [9] en effectuant toutes les combinaisons possibles entre les arguments, chaque argument conservant toutefois son rang. La structure hiérarchique partielle est ainsi enrichie de plusieurs dépendances simples pour chaque dépendance disjonctive – qu'elles remplacent. La correspondance des structures de la question et de la réponse candidate peut alors être une nouvelle fois testée.

Le succès de la comparaison des structures déclenche la sélection de la réponse candidate et sa présentation à l'utilisateur comme réponse pertinente à la question posée. Le système passe ensuite à la réponse candidate suivante pour lui appliquer la même procédure si une autre candidate est proposée, ou bien s'arrête.

Mais s'il n'y a pas eu correspondance entre la structure locale de la question et celle de la réponse candidate, il est inutile d'avoir recours une nouvelle fois à la structure syntaxico-sémantique de la base textuelle pour y puiser une

[7] Excepté en ce qui concerne l'argument des dépendances impliquant l'objet de la question. En effet, un argument de ces dépendances ne présente aucune unité lexicale, mais bien une ou plusieurs exigences liées à des traits sémantiques.

[8] Celle de la requête doit être entièrement incluse dans celle de la réponse candidate.

[9] Par exemple, une dépendance disjonctive de type `SUBJ` donnera plusieurs dépendances de type `SUBJ`, une dépendance disjonctive de type `VARG[DIR]` donnera une série de dépendances de type `VARG[DIR]`, etc.

information supplémentaire à verser dans cette structure candidate. En effet, l'ensemble des données disponibles y sont maintenant présentes. Pourtant, il est possible que la réponse candidate apporte une réponse pertinente à la question, mais que les contraintes imposées pour mettre en correspondance réponses et question soient trop strictes pour que cette réponse convienne.

Deux possibilités de relâchement des contraintes se présentent alors :

- il est possible de réduire la quantité d'information présente dans la structure informationnelle de la question et, par voie de conséquence, d'élargir les possibilités de correspondance entre cette structure et celle des réponses candidates [10]. Cependant, nous sommes engagés ici dans une procédure qui réclame une complète correspondance entre les structures plates de la requête et des réponses candidates. Cette technique est donc à rejeter pour l'instant [11] ;
- il est également possible de se contenter de correspondances partielles entre les dépendances contenues dans la structure hiérarchique de la question et la structure locale hiérarchique de la réponse candidate [12]. C'est cette méthode de relâchement des contraintes sur la correspondance que nous utilisons à ce stade de la méthode.

L'application du relâchement sur la correspondance des structures de la requête et de la réponse candidate provoque l'attribution d'un score à cette réponse candidate qui correspond à la proportion des dépendances de la question présentes dans la réponse candidate. Les réponses candidates qui coïncident partiellement avec la structure de la requête sont conservées en mémoire et peuvent être classifiées en fin de traitement de la question en fonction du score qu'elles obtiennent. Du fait de leur plus faible correspondance avec la question, le niveau de confiance qui leur est attribué est susceptible de varier en fonction du score qui leur est attribué. Il est également possible d'éliminer certaines réponses candidates si leur score n'est pas suffisamment élevé. Le niveau du score d'élimination est paramétrable.

6.3.2 Diminution de l'information de la requête.

Au cours de notre examen d'un corpus de questions (*cf.* annexe B page 279) posées à TREC (en anglais), nous avons noté que souvent l'unité lexicale qui détermine l'objet de la question à l'intérieur de la requête est soit un hyponyme, soit un hypéronyme du terme qui sera réellement utilisé dans le texte.

[10] Nous appelons « **dégradation sur la question** » ce type de relâchement des contraintes.

[11] La section suivante 6.3.2 étudie l'opportunité de limiter l'information requise pour mettre en correspondance le contenu de la question et le contenu de fragments de textes appelés à devenir des réponses candidates.

[12] Nous appelons « **dégradation sur la correspondance** » ce type de relâchement des contraintes.

Nous avons signalé que le dictionnaire *EuroWordNet* contient une taxinomie hyponymique pour un certain nombre d'unités lexicales, mais que nous n'avons pu mettre en œuvre son intégration dans la structure informationnelle (*cf.* section 4.3.2 page 148).

Ce défaut de taxinomie hyponymique qui permettrait de construire un lien entre ce terme catégorisant l'objet de la question et le terme qui constitue son pendant dans le texte empêche une mise en correspondance des structures informationnelles de la question et de fragments de textes qui devraient être identifiés comme des réponses candidates. C'est également dans la perspective de combler cette lacune que nous avons imaginé la dépendance `FOCUS` qui a pour vocation d'identifier l'objet de la question.

En effet, nous exploitons ici l'identification de ce *focus* pour effectuer une diminution de la contrainte lexicale dans la construction de la structure informationnelle liée à la question. Ainsi, au même titre que l'interrogatif est éliminé des dépendances extraites de la question pour être remplacé, le cas échéant, par les traits sémantiques qui le caractérisent, l'unité lexicale correspondant au *focus* sera éliminée des dépendances dans lesquelles elle apparaît pour être remplacée par les traits sémantiques qui lui sont propres et qui constituent de ce fait une contrainte plus souple que l'apparition d'un lexème dans la structure informationnelle de la question.

Une fois cette opération effectuée, la méthode reste la même que pour la mise en correspondance des questions et des réponses avec des structures plates compatibles dans leur totalité. La procédure sélectionne d'abord des réponses candidates à partir des structures plates. Ensuite, elle vérifie leur correspondance avec la structure hiérarchique au niveau des dépendances simples, puis éventuellement au niveau des dépendances disjonctives. Si la réponse n'est pas confirmée à ce stade du traitement, on peut de nouveau relâcher les contraintes sur la correspondance et donner un score de coïncidence à la réponse candidate en fonction du nombre de ses dépendances qui correspondent parfaitement avec celles de la question.

Il faut noter que les relâchements de contraintes sont paramétrables par l'utilisateur, qui peut à sa guise demander l'un ou l'autre de ces assouplissements d'exigences et, dans le cas de l'attribution d'un score, fixer la mise en correspondance des structures à un certain niveau de confiance. Il est également possible d'approfondir le relâchement sur la question en supprimant d'autres unités lexicales des dépendances où elles apparaissent pour les remplacer par des contraintes sur leurs traits sémantiques. Nous n'avons pas progressé plus avant dans cette direction qui intuitivement suggère une génération importante de bruit, mais il est probable que le nombre de réponses correctes identifiées par le système s'en ressentirait favorablement. Toutefois, le fait qu'aucune désambiguïsation sémantique ne soit appliquée au niveau

de la question autorise rapidement une très large augmentation des réponses candidates et sans doute des réponses jugées pertinentes, parfois à tort.

Une deuxième remarque concernant les capacités inexploitées de ce système porte sur la possibilité de placer des pondérations plus ou moins importantes sur les dépendances. En effet, toutes les dépendances syntaxiques ne possèdent pas le même pouvoir expressif, c'est-à-dire qu'elles n'expriment pas un sens avec la même intensité. Nous avons d'ailleurs éliminé les dépendances fonctionnelles au motif qu'elles n'avaient de rôle que syntaxique, et non syntaxico-sémantique ou sémantique. Il est donc possible, et même probable que certaines dépendances aient plus d'importance que d'autres dans la détermination de la validité d'une réponse candidate. Ces pondérations pourraient être utilisées notamment dans l'exploitation du relâchement sur la correspondance. Nous n'avons pas testé cette fonctionnalité qui s'éloigne des méthodes linguistiques classiques mais présente un intérêt réel dans le cadre d'une application fonctionnelle.

Il aurait enfin été intéressant d'effectuer un traitement de la question entièrement analogue à celui des documents de la base textuelle, et de profiter à ce niveau également de l'enrichissement. Nous avons parlé déjà des inconvénients et des dangers de ce choix. Nous renouvelons nos réticences tout en déplorant de n'avoir pu réaliser d'essai réel en ce sens.

Avec les traitements appliqués aux requêtes, tantôt parallèles à ceux des document et tantôt très différents, puis avec la procédure de mise en correspondance de ces requêtes avec des fragments de textes susceptibles d'y apporter une réponse, nous avons non seulement établi une méthode permettant de valoriser l'information contenue dans la structure syntaxico-sémantique construite à partir de la base textuelle, mais nous avons également fourni un outil d'interrogation largement paramétrable qui permet de répondre à des besoins très divers des utilisateurs. L'application de question-réponse en est une illustration, mais l'extraction d'information est également possible, et son interrogation au niveau du document permet de l'associer au filtrage de textes. Il nous reste à présent à évaluer les différents processus qui d'une part construisent la structure informationnelle et d'autre part se chargent de son interrogation plus ou moins stricte.

6.4 Conclusion

Cette interrogation en langage naturel rend possible le questionnement du système comme on le ferait pour un être humain. Une analyse lexico-syntaxique semblable à celle qui a extrait l'information de la base documentaire permet de construire la structure de base de la question et d'en identifier l'information. En particulier, une grammaire spécifique autorise dans

une certaine mesure la détermination de l'objet de cette question (*focus*).

La réponse à la question posée consiste en un fragment de texte de la base documentaire (principalement la phrase, mais des besoins différents peuvent y préférer le paragraphe ou le document) qui contient, originellement ou suite aux enrichissements, l'ensemble des éléments informatifs identifiés dans la question, y compris un type correspondant à l'objet de la question. Une réponse idéale permet d'apparier totalement la structure de la réponse à celle de la question. Cependant, la formulation de la question et celle de la réponse peuvent diverger à un point tel que l'enrichissement ne suffit pas pour établir une correspondance totale. Dès lors, un appariement dégradé de la réponse avec la question est possible.

Il reste maintenant à effectuer des tests en interrogeant une base documentaire sur une échelle suffisante pour pouvoir évaluer les points forts et les faiblesses de notre méthodologie.

Chapitre 7

Évaluation de la méthode d'enrichissement

7.1 Introduction

À partir d'outils d'analyse et de ressources d'ordre linguistique, nous avons conçu un système complet qui permet d'exploiter l'information contenue dans une base de textes en français. Ce système comporte deux aspects :

- l'élaboration de la structure informationnelle qui s'effectue en deux phases : l'analyse des textes et l'extraction de l'information qu'ils contiennent pour constituer une structure syntaxico-sémantique, puis l'enrichissement de cette structure. C'est principalement sur cet aspect que notre travail a porté ;
- l'interrogation de la structure informationnelle par des requêtes en langage naturel, basée sur une analyse de même nature que celle des textes et aidée par certains traitements particuliers. Cette partie du système a été entièrement élaborée au cours de notre recherche.

Il s'agit à présent d'estimer la qualité non seulement du fonctionnement complet du système, mais aussi des traitements particuliers qui sont mis en œuvre au cours du fonctionnement de l'application. Notamment, il s'agit d'évaluer l'intérêt propre à chacune des méthodes utilisées pour recueillir les différentes informations qui sont emmagasinées :

- l'analyse morpho-syntaxique réalisée par les outils d'analyse *NTM* et *XIP* développés à XRCE ;
- la désambiguïsation sémantique basée sur la méthode de XRCE que nous avons réimplantée et adaptée à nos besoins ;
- les différents types d'enrichissements lexicaux, syntaxiques et sémantiques tels que nous les avons définis.

Par ailleurs, le système d'interrogation de la structure informationnelle a soulevé plusieurs questions qui nous ont amené à nous intéresser à l'identification de l'objet de la question et à la détermination d'un seuil de confiance en la réponse proposée lorsque l'information contenue dans la requête ne correspond qu'imparfaitement à celle de cette réponse. Nous avons donc autorisé un paramétrage de ces deux données afin de pouvoir vérifier leur impact sur les résultats obtenus. Ce paramétrage agit sur le niveau des contraintes appliquées à la correspondance entre question et réponses candidates. Il existe deux critères qui peuvent varier :

- variation de la proportion des dépendances correspondantes entre requête et réponses candidates ;
- abstraction sémantique de l'objet de la question contenu dans la requête pour faciliter la correspondance entre requête et réponses candidates.

Enfin, la méthodologie que nous proposons ne se limite pas à la tâche de question-réponse, qui est notre moyen de tester les performances des divers processus qui composent le système. En effet, la structuration de l'information qui en constitue la clef de voûte a été conçue pour permettre une gestion des données contenues dans la base textuelle à diverses fins, et par exemple le filtrage de textes ou l'extraction d'information. Dans le but d'autoriser ces diverses manipulations, nous avons envisagé de traiter la structure informationnelle à différents niveaux de découpage des textes : dépendance, phrase, paragraphe, document.

Le présent chapitre va s'attacher à établir un protocole d'évaluation de la méthode de construction de la structure informationnelle d'une base textuelle, et donc de la qualité de cette base. Au cours de cette évaluation, la structure sera interrogée au moyen du système d'interrogation de la base documentaire que nous avons développé. Le processus d'évaluation portera donc à la fois sur les deux méthodes que nous avons proposées. D'une part, la méthode de constitution de la structure informationnelle sera examinée à travers les différents traitements qui apportent une information à cette structure. D'autre part, la procédure d'interrogation de la structure sera testée à chaque niveau de contraintes qu'elle autorise. Nous ferons ensuite une présentation critique des résultats obtenus.

7.2 Définition des critères

Pour définir une stratégie d'évaluation qui décrive bien les qualités et les défauts d'une méthodologie, il faut avant tout identifier les objectifs et les capacités théoriques de cette méthodologie et en tenir compte pour ne pas établir des critères d'évaluation qui sortent de son domaine. En effet, il n'est pas raisonnable de juger la capacité d'une application à résoudre un

problème auquel cette application n'est pas destinée.

7.2.1 Objectifs et aptitudes de notre méthode

La méthodologie proposée ici permet de gérer une requête en langage naturel à l'aide de méthodes linguistiques pour retrouver dans une base textuelle, analysée et traitée par des outils et des ressources tout aussi linguistiques, le ou les fragments de texte qui en constituent la réponse. Sa vocation est donc d'apporter une réponse aux questions ouvertes [1] qui lui sont proposées.

L'utilisation exclusive de méthodes linguistiques et le fait que seuls des fragments de texte existants constituent les réponses interdisent en effet les jugements de valeur de vérité sur la question, qui demandent généralement des réponses positives ou négatives. Pour les mêmes raisons, les réponses explicatives – par exemple à la plupart des questions introduites par « pourquoi » ou « comment » – ne sont pas envisageables.

Par ailleurs, la méthode que nous avons mise en place comporte deux lacunes par rapport aux traitements que nous avions initialement prévu de lui assigner. En effet, deux types de traitements n'ont pu être intégrés dans l'application faute de temps, et la structure informationnelle ne peut dès lors bénéficier de leurs apports :

- le premier module devait gérer les problèmes de coréférence des pronoms personnels et des pronoms et adjectifs possessifs. Le prototype de [Trouilleux, 2001] constitué essentiellement d'une grammaire *XIP* avait été envisagé pour effectuer cette opération mais il demandait un travail d'adaptation important notamment du fait de son analyse effectuée au niveau du texte et non au niveau de la phrase ;
- le second de ces modules est chargé de gérer l'implication logique. Nous comptions exploiter pour cela les différents types d'implication logique présents dans *EuroWordNet* (coexistence, inclusion, présupposition et cause, *cf.* section 3.4.3 page 120 et [Vossen, 1997]). Comme nous l'avons dit (*cf.* section 4.3.2 page 148), le temps nous a manqué pour exploiter l'ensemble de l'information présente dans *EuroWordNet*. Du fait de la relative limitation du lexique traité dans cette ressource, nous avons privilégié les démarches plus générales.

Mener l'évaluation d'un système qui interroge de l'information d'une base documentaire sans avoir résolu, même partiellement, le problème de la coréférence d'entités n'aurait pas de sens, vu la fréquence de ce type de construction syntaxique dans les textes. Dans le type de texte que nous avons à interroger – un dictionnaire encyclopédique – c'est plus que jamais le cas car la plupart

[1] C'est-à-dire des questions auxquelles on ne peut répondre par *oui* ou par *non*, par opposition aux questions fermées.

du temps, le mot-vedette de l'article n'est pas explicitement nommé dans le corps du texte : c'est un simple pronom qui y fait référence. Puisque la partie de la grammaire qui permettrait de gérer le problème de la coréférence n'a pu être intégrée à notre système faute de temps pour l'adapter, il a fallu nous résoudre à utiliser une simulation naïve de la coréférence, à l'efficacité relative, mais rapide à mettre en œuvre et permettant d'éliminer dans certains cas le problème visé.

La méthode que nous avons choisie ne pourra fonctionner pour tous les types de texte, mais elle ne se limite pas à un corpus d'évaluation car c'est une méthode qui se prête parfaitement à une structure de dictionnaire. En effet, le thème traité dans chaque article correspond à l'intitulé de cet article, parfaitement défini. C'est à cet intitulé que la plupart des **anaphores** contenues dans l'article font référence. La méthode consiste donc, pour toute dépendance de type `SUBJ` dont le second argument (c'est-à-dire le sujet du verbe qui constitue le premier argument) est un pronom personnel à la troisième personne, à construire une dépendance disjonctive. Dans cette dépendance, la disjonction porte sur l'argument sujet, et met en disjonction le pronom personnel présent dans la phrase et la vedette de l'article qui en constitue le thème.

Afin d'asseoir cette méthode, nous avons étudié environ trois cents articles de l'*Encyclopédie* pour constater que dans la plupart des cas, les sujets pronominaux renvoyaient à l'intitulé de cet article. C'est particulièrement sensible dans les sujets qui concernent des personnages. Bien entendu, ce n'est pas une règle absolue et dans certains cas (plus de 30 %), l'identification du sujet pronominal à l'intitulé de l'article n'est pas correcte. Toutefois, l'introduction d'une dépendance disjonctive permet de ne pas perdre l'information de la dépendance initiale même dans les cas où une coréférence erronée s'est introduite dans la structure sous la forme d'une disjonction. De plus, l'assignation d'un trait `coref` à l'argument ajouté donne la possibilité de ne pas exploiter cette information lors de l'interrogation si elle est considérée comme trop peu sûre.

Par contre, nous n'avons pas trouvé de méthode rapide à mettre en œuvre pour gérer l'inférence logique. De ce fait, notre méthode est actuellement incapable de manier l'induction, la déduction ou toute autre opération d'ordre logique. Dès lors, l'information qu'elle est apte à détecter et manipuler doit être explicite dans la base documentaire, sans s'appuyer sur aucun artifice d'ordre logique. Par contre, toutes les formes linguistiques possibles de chaque énoncé sont envisageables dans le cadre de la grammaire du français.

7.2.2 Les campagnes d'évaluation en gestion de l'information

Si la définition d'un protocole d'évaluation doit tenir compte des caractéristiques de l'application à tester, elle se doit également de respecter une réelle indépendance par rapport à cette application pour se garder de définir une expérience adaptée aux capacités matérielles de l'application, et non à ses objectifs méthodologiques. Pour éviter cet écueil, nous nous sommes intéressé aux campagnes d'évaluation menées dans le domaine. Essentiellement, nous avons étudié l'approche mise en œuvre lors des compétitions TREC (*Text REtrieval Conference*), ouvertes aux systèmes de question-réponse pour l'anglais depuis sa huitième édition [Voorhees et Harman, 1999b].

Toutefois, les critères d'évaluation de cette compétition évoluent à chaque édition, et tant la conception que le déroulement de cette évaluation sont de plus en plus exigeants en temps et en moyens. Dans la limite de nos possibilités, nous avons donc étudié le mode de fonctionnement de l'évaluation menée dans le cadre de TREC-8 [Voorhees, 1999].

Cette première version d'une évaluation de systèmes de question-réponse était effectuée à l'aide d'un ensemble de 198 questions factuelles appelant des réponses courtes extraites d'une base textuelle ou reconstruites à partir de cette base. Chacune des questions possédait au moins une réponse pertinente dans les textes. En effet, ces questions ont été générées par des opérateurs humains à partir des documents de la base textuelle.

Les résultats attendus sont une liste de 5 paires `[numéro de texte, chaîne de caractères-réponse]` maximum ordonnées en fonction d'un score. Ce score correspond au degré de certitude que la réponse attendu est bien contenue dans le fragment de texte. Les réponses sont limitées soit à 250 caractères, soit à 50 caractères et peuvent constituer un fragment de texte issu de la base textuelle ou être générés à partir de l'information contenue dans la base textuelle.

L'évaluation de ces réponses est dichotomique [2] et réalisée par des experts humains. La quantification s'effectue par attribution à chaque question d'un score égal au nombre inverse du rang auquel apparaît la première bonne réponse si cette réponse se trouve dans les cinq premières fournies, et 0 sinon. Ainsi, une première bonne réponse au premier rang sera créditée d'un score de 1 ($\frac{1}{1}$), au deuxième rang d'un score de 0,5 ($\frac{1}{2}$), etc. Les questions n'obtenant pas de bonne réponse ou celles qui n'obtiennent pas de réponse du tout ont un score nul. Le fait de donner plusieurs bonnes réponses n'intervient pas sur le score accordé à la question [Voorhees et Harman, 1999a].

[2] C'est-à-dire que chaque réponse est soit correcte, soit incorrecte, sans gradation de qualité.

Les éditions ultérieures de TREC vont modifier plus ou moins ces critères. Ainsi, TREC-9 [Voorhees, 2000] constitue un ensemble de 500 questions à partir de requêtes réellement posées sur l'encyclopédie *Encarta*, et construit un corpus de textes à partir de ces questions. Cependant, chaque question possède encore au moins une fois sa réponse dans la base textuelle, ce qui n'est plus le cas lors de la conférence TREC-10 [Voorhees, 2001]. Cette différence modifie les mesures d'évaluation puisqu'il est correct de ne pas rendre de réponse lorsque la question n'a pas de réponse dans la base textuelle. Cette réponse vide reçoit un score de la même manière que les autres réponses, en fonction de son rang dans la liste. Lors de cette dixième compétition, la réponse de 250 caractères est supprimée et seules des réponses de 50 caractères sont acceptées.

Enfin, la onzième édition de TREC, qui s'est tenue en novembre 2002, modifie le calcul de l'évaluation d'une part en limitant à une seule les réponses attendues pour chaque question, d'autre part en réclamant la réponse réelle uniquement, et non plus une fenêtre déterminée par un nombre de caractères [Voorhees, 2002]. Nous n'avons pu tenir compte des caractéristiques de cette dernière conférence, qui a eu lieu alors que nous avions mis en place nos propres critères d'évaluation.

Conçues sur le modèle de TREC, deux autres conférences destinées à l'évaluation de systèmes de gestion de l'information ont vu le jour. Il s'agit d'abord de la conférence **CLEF** (*Cross-Lingual Information Forum*) qui intègre plusieurs langues européennes (anglais, allemand, français, italien, espagnol, néerlandais, suédois, finnois) [Peters, 2002], mais cette campagne d'évaluation n'intègre la tâche de question-réponse que dans son édition de 2003. L'autre conférence qui imite TREC s'appelle **NTCIR** (*NII-NACSIS Test Collection for IR Systems*) et est destinée aux langues asiatiques (japonais, coréen, chinois). La tâche de question-réponse y est définie par le seul système qui participe à la campagne d'évaluation pour cette discipline [Fukumoto et Kato, 2001, Fukumoto et al., 2003] et est comparable à celle de TREC-8.

7.2.3 Définition de nos critères d'évaluation

Un cas de figure idéal nous verrait utiliser un protocole d'évaluation existant et reconnu par la communauté du domaine pour tester la qualité de notre approche. Toutefois, cette situation ne peut se présenter car aucune évaluation n'a jusqu'à présent obtenu de large consensus dans le domaine des applications de question-réponse en langue française. Par ailleurs, l'adaptation d'un protocole d'évaluation existant ne peut se faire qu'en considérant le manque relatif des moyens dont nous disposions pour la réaliser. En effet, nous n'avons pu obtenir les questions posées par les utilisateurs de la

version en ligne de l'*Encyclopédie Hachette Multimédia*,[3]. De ce fait, il nous est impossible de construire une base textuelle nécessaire à l'expérience à partir d'un corpus de questions, comme c'est le cas à partir de la neuvième compétition TREC. D'autre part, il n'est pas réaliste de mener seul et dans le temps qui nous est imparti une évaluation de la même ampleur que celles qui sont proposées dans des structures telles que TREC. Un corpus de 200 questions nous semble raisonnable et déjà significatif.

Ces différentes considérations nous ont amené à préférer le mode d'évaluation choisi lors de la huitième conférence TREC [Voorhees, 1999] dont la procédure ne dépasse pas les 200 questions et qui constitue ce corpus de requêtes à partir des textes de la base documentaire assemblée *a priori*. Chacune de ces questions trouve au moins une fois sa réponse dans la base documentaire.

Toutefois, comme ce protocole évalue des applications qui traitent des textes en anglais, il nous est impossible de reproduire complètement l'expérience. De plus, notre méthode de structuration sémantique de l'information textuelle n'a pas pour seule vocation une application de question-réponse, mais cherche à faciliter tous les besoins de gestion de l'information. Il s'agit donc d'intégrer dans son évaluation des tests permettant de juger son efficacité dans d'autres secteurs du domaine, et notamment la mesure du rappel.

Constitution d'un corpus

La constitution d'un corpus a répondu à ces besoins. Nous avons rassemblé un ensemble de 50 articles de l'*Encyclopédie Hachette Multimédia* représentant environ 20 000 mots. Ces articles sont de taille variable, les plus courts ne contenant que deux ou trois phrases (74 mots, 2 Ko) et les plus longs plusieurs paragraphes (2 576 mots, 25Ko). La taille réduite de ce corpus s'explique par le fait que nous devons en maîtriser entièrement l'information pour pouvoir juger du rappel du système et pour évaluer sa qualité dans des applications autres que la tâche de question-réponse. Dans ce but également, nous avons constitué ce corpus dans un domaine restreint et précis afin de pouvoir disposer d'une information qui se recoupe dans plusieurs articles. Cette information redondante permet de juger du rappel du système, et met également sa précision à l'épreuve, car ce type de corpus favorise une certaine confusion.

Le domaine que nous avons choisi pour construire ce corpus est celui des personnalités romaines de l'Antiquité. Le choix d'articles portant sur

[3] L'*Encyclopédie Hachette Multimédia* est consultable gratuitement sur différents sites Internet : http://www.encyclopedie-hachette.com/W3E/, http://www.club-internet.fr/encyclopedie/, http://encyclo.voila.fr/ ou http://fr.encyclopedia.yahoo.com/.

des personnages découle de la méthode de coréférence analogique que nous avons intégrée dans le traitement des textes. Il ne s'agit pas ici de tester cette méthode, dont nous avons dit la naïveté, mais plutôt de juger de ce qu'une méthodologie plus subtile pourrait apporter. Nous avons constaté que les articles portant sur des personnes physiques sont ceux qui bénéficient le plus de la technique que nous employons. C'est donc dans ce cadre que cette technique se rapproche au maximum des résultats que pourrait présenter une approche plus fiable et raffinée. Aussi ces articles nous permettront-ils de tirer des conclusions sur l'apport de la résolution des coréférences sur les performances de notre système.

Pour le reste, la construction du corpus de texte s'est effectuée automatiquement, sur la base des dates de naissance et de mort des personnages et de la caractéristique *romain* (ou *romaine*) signalée dans la balise XML `Resume` qui énonce brièvement les principales caractéristiques du sujet de l'article (*cf.* ce champ dans la figure 1 page 23). Les articles du corpus sont sélectionnés au hasard dans la liste classée par ordre de taille des documents répondant aux spécifications.

Génération des questions

La création des questions a été confiée à huit opérateurs qui n'ont pas participé à l'élaboration des outils que nous utilisons dans notre système, et qui ne sont que peu ou pas du tout informés de notre méthodologie. Nous avons de plus essayé de constituer deux ensembles de testeurs différents en choisissant quatre d'entre eux dans le domaine de la linguistique computationnelle mais en dehors du cadre de la gestion de l'information[4] et les quatre autres dans des domaines qui ne touchent ni à la linguistique ni à la recherche d'information.

Nous leur avons donné à chacun une partie du corpus correspondant à un huitième du total des textes en leur demandant d'en étudier le contenu pour poser des questions sur l'information qu'ils comportent. Ces questions doivent répondre à certaines exigences. Tout d'abord, elles doivent concerner un élément *explicite* du texte. La réponse doit donc être présente et ne pas dépendre d'une déduction logique basée sur le contenu du document. Par exemple, le système ne peut savoir que Jules César est défunt en 43 a.C.n.[5] que si cette information est formulée dans le texte, et non parce qu'il est indiqué qu'il est assassiné en 44 a.C.n. La réponse ne peut pas non plus porter

[4] Ces testeurs travaillent dans le cadre de l'analyse ou de la génération de textes en langage naturel, et non dans des domaines comme la recherche d'information, le filtrage, l'extraction d'information, la tâche de question-réponse, etc.

[5] Il n'existe aucun consensus dans l'abréviation qui indique les dates avant notre ère et après elle. Nous utilisons l'abréviation latine *a.C.n.* pour *ante Christi natum* (avant la naissance du Christ) et *p.C.n.* pour *post Christi natum* (après la naissance du Christ).

un jugement sur l'énoncé de la question. Ainsi, des questions appelant une réponse positive ou négative (« Est-ce que... »), ainsi que celles qui peuvent attendre une explication (« Pourquoi... », « Comment... ») sont également proscrites, car on ne peut généralement trouver un extrait de texte qui leur répond.

Par ailleurs, comme notre système ne gère pas les listes – qui ne sont pas prises en compte non plus par la compétition TREC-8 – les réponses attendues doivent être élémentaires. Par contre, toutes les variations d'ordre linguistique sur l'énoncé de la question sont possibles. Nous avons demandé un minimum de 25 questions par testeur et un maximum de 28. Les trois questions surnuméraires permettent de remplacer les éventuelles transgressions des principes de création des questions. Si aucune des questions ne transgresse ces principes, ce sont les 25 premières qui sont conservées.

Une fois l'ensemble des questions rassemblées, nous les avons parcourues afin d'en corriger les éventuelles erreurs orthographiques ou grammaticales [6], et surtout nous avons vérifié le respect des critères que nous avions définis. Sur les 206 questions proposées par nos testeurs, nous en avons éliminé six dont plusieurs des règles de génération n'avaient pas été respectées. Nous avons donc conservé les 200 questions qui constituent la base de test de notre évaluation. Les testeurs proposaient une réponse par question posée.

Nous avons ensuite dû parcourir l'ensemble du corpus pour rechercher les autres possibilités de réponse pour chaque question. Nous avons donc obtenu pour chacune de ces requêtes une liste de coordonnées permettant de situer la réponse dans les documents. Cette liste permet de juger de la qualité des réponses données par le système, mais aussi de son rappel en comptabilisant les réponses exactes données par rapport aux questions attendues.

La fenêtre de réponse que nous utilisons ne correspond pas aux spécifications de TREC-8, qui prévoyaient un champ de 50 caractères ou de 250 caractères. En effet, ces fenêtres sont totalement arbitraires et ne correspondent à aucun découpage réaliste de l'information, qu'il soit syntaxique ou sémantique. La phrase est un élément unitaire de réponse plus pertinent. Elle est assez proche de la fenêtre de 250 caractères lorsqu'elle est longue, de celle de 50 caractères lorsqu'elle est courte [7]. Moyennant des modifications importantes dans le module qui permet d'extraire du document le groupe syntaxique désiré, il est possible de réduire la fenêtre de réponse à un syntagme ou à une suite de syntagmes. La principale difficulté de cette

[6] En effet, notre système, n'étant pas pourvu d'un système de correction orthographique, suppose des données d'entrée parfaitement orthographiées.

[7] La conférence TREC-11 [Voorhees, 2002] a supprimé une valeur absolue de fenêtre au bénéfice de la réponse exacte à la question. Les critères d'évaluation de cette compétition ont été diffusés en novembre 2002, alors que notre application et notre protocole d'évaluation étaient finalisés.

modification vient de la reconstruction du syntagme à partir d'un groupe syntaxique partiel construit par l'analyseur syntaxique *XIP*.

Modes de calcul des résultats

Enfin, le calcul des résultats de cette évaluation présente deux aspects. D'abord, ils seront calculés conformément aux spécifications de TREC-8 dont nous nous sommes inspiré du mode opératoire. De ce fait, à chaque question est associée une liste de cinq réponses maximum [8], qui sont ordonnées selon leur degré de pertinence, ou selon le degré de confiance en leur validité. Dans notre système, ce degré de confiance correspond aux taux de coïncidence entre l'information des réponses retenues et l'information contenue dans la question. Chacune des réponses de la liste reçoit un score égal au nombre inverse du rang de classification de la première bonne réponse [9], et l'absence de bonne réponse aboutit à un score nul.

Un autre mode de calcul des résultats s'éloigne de la tâche de question-réponse pour se rapprocher des autres secteurs de la gestion de l'information. Dans ce mode de calcul des résultats, les réponses correctes, incorrectes et manquantes sont prises en compte au travers des mesures de précision et de rappel. La F-mesure propose une moyenne entre les résultats de précision et de rappel donnant la préférence tantôt à la précision ($\beta = 0{,}5$), tantôt au rappel ($\beta = 2$), tantôt en leur conférant une importance équivalente ($\beta = 1$) (*cf.* section 1.2.1 page 26). Lorsque nous calculons ce type de résultats, nous fournissons également au niveau quantitatif le nombre de réponses exactes, le nombre de questions qui obtiennent au moins une réponse exacte et le nombre de réponses fausses.

Ce second mode de calcul nous amène à nous pencher sur la notion de fenêtre que nous avons abordée précédemment. En effet, le choix d'une fenêtre de réponse d'une phrase convient bien à la tâche de question-réponse, et éventuellement à celle d'extraction d'information. Mais d'autres applications peuvent chercher à exploiter l'information à d'autres niveaux. Par exemple, le filtrage de textes travaille au niveau du document. Nous avons déjà signalé que nous emmagasinons l'information à différents niveaux de profondeur : texte, paragraphe, phrase, dépendance. Ce dernier niveau n'est pas opérationnel actuellement, mais les autres niveaux sont parfaitement exploitables. Nous effectuons donc le calcul des résultats selon le second mode d'évaluation aux différents niveaux de stockage de l'information, ce qui implique également divers niveaux d'interrogation.

[8] Sauf dans le cas où aucune réponse n'est trouvée.

[9] Si la première bonne réponse de la liste est au cinquième rang, la question recevra un score de $\frac{1}{5}$, si elle est au quatrième rang, le score sera de $\frac{1}{4}$, etc.

Pour l'ensemble des mesures que nous calculons, que ce soit pour la tâche de question-réponse ou les autres secteurs de la gestion de l'information, nous établissons des résultats pour chaque variation des paramètres qui nous sont permis. Nous pouvons notamment faire varier :

- l'utilisation ou non de l'unité lexicale désignée par la dépendance `FOCUS` dans la question ;
- l'utilisation ou non d'un seuil (variable) de correspondance des dépendances entre la question et la réponse ;
- l'utilisation ou non de la méthode naïve de coréférence des pronoms sujets ;
- l'utilisation et la combinaison des méthodes d'enrichissement et des ressources qui fournissent leur information [10].

Un niveau d'efficacité « **plancher** » [11] (*baseline*) a été défini pour les deux types de mesure. Ce plancher est calculé de la même manière que les autres résultats, mais à partir d'un corpus sans enrichissement ni analyse. Les mots de la question (substantifs, verbes et adjectifs) sont utilisés comme des mots-clefs pour rechercher la réponse dans les documents qui n'ont reçu aucun des traitements de notre approche. Pour juger l'efficacité d'un enrichissement synonymique contextuel par rapport à une expansion synonymique aveugle, un deuxième niveau plancher [12] est calculé où tous les mots sont enrichis avec toutes les possibilités de synonymie proposées par le dictionnaire *Dubois*.

7.3 Présentation des résultats

7.3.1 Examen des résultats selon les critères de question-réponse

L'examen des résultats de l'évaluation va nous permettre d'identifier les points forts et les faiblesses de notre méthode. La première démarche à envisager est l'examen global de notre méthodologie, qui valide ou infirme l'utilité d'un traitement basé sur des méthodes linguistiques dans le cadre de la gestion de l'information. C'est d'abord dans la perspective de la tâche de question-réponse que nous testons notre système. Pour ce faire, nous confrontons les mesures de plancher avec les résultats de la méthode dite « **globale** ».

[10] L'ensemble des tableaux de résultats est fourni en annexe C page 281.

[11] Ce niveau d'efficacité plancher est étiqueté « plancher » dans les tableaux de présentation des résultats.

[12] Lorsque nous parlons des planchers (au pluriel), il s'agit de ces deux structures qui n'exploitent rien de notre méthodologie.

Analyse des résultats de la méthode globale

Ce que nous appelons méthode globale est une application du système sur les textes qui met en œuvre l'ensemble des processus et enrichissements disponibles (analyse morpho-syntaxique, désambiguïsation sémantique, résolution de la coréférence des pronoms sujets, synonymie, traits sémantiques, dérivation morphologique). Au niveau des contraintes paramétrables, cette méthode globale privilégie d'une part la précision en imposant la présence de l'unité lexicale *focus* (*cf.* section 6.2.2 page 187), mais sans négliger le rappel, car la mise en correspondance des dépendances de la requête et des réponses candidates ne rejette ces dernières que si aucune coïncidence n'est constatée.

Le tableau 7.1 indique clairement que notre méthode améliore non seulement la qualité des réponses apportées en promouvant plus efficacement les bonnes réponses (amélioration : 45 %), mais augmente également son potentiel à apporter une réponse à la question posée (amélioration : 19 %). L'enrichissement synonymique aveugle n'apporte que peu d'amélioration par rapport au plancher. Toutefois, il est difficile de juger de ce résultat sans disposer des chiffres qui évaluent un enrichissement, certes semblable à celui-ci, mais soumis à la sélection de la désambiguïsation sémantique. Ce tableau ne permet pas non plus d'identifier l'apport respectif des enrichissements qui nous ont conduit à ces résultats.

Enrichissement	Score	Sans réponse
Plancher	0,295	139
Syn. aveugles	0,303	137
Tous types	0,428	113

TAB. 7.1 – Résultats des planchers et de la méthode globale selon le protocole de question-réponse.

La table 7.2 page 210 présente donc les résultats comparés des différents modes d'enrichissement utilisés dans le cadre de la méthode globale. La présentation du tableau montre un enrichissement croissant :

- le premier de ces enrichissements (***Synonymie aveugle***) a déjà été présenté : il s'agit du plancher enrichi par les synonymes du *Dubois* mais sans discernement. Aucune analyse n'a été apportée et aucune résolution de coréférence n'a été réalisée. Le reste du tableau par contre correspond à la méthode globale et la méthode de coréférence a été utilisée ;
- la ligne ***Syntaxe*** présente les résultats de l'interrogation après analyse morpho-syntaxique du texte. Les traits sémantiques sont ajoutés à ce stade et les enrichissements suivants profitent de l'analyse syntaxique ;

- le champ ***Synonymes*** indique l'apport d'une synonymie aveugle appliquée après analyse syntaxique des documents ;
- la ligne ***Dubois*** (***D***) indique un enrichissement réalisé à l'aide des synonymes contextuels du dictionnaire *Dubois* suite à la désambiguïsation sémantique. À partir de cette ligne, tous les enrichissements bénéficient de cette désambiguïsation ;
- la ligne ***Dubois - Synonymes*** (***D-Syn***) met en œuvre l'enrichissement contextualisé pour les synonymes du dictionnaire *Dubois*, mais applique également un enrichissement aveugle des mots qui n'ont pas été désambiguïsés à partir de ces mêmes synonymes issus du *Dubois* ;
- la ligne ***Dubois - Bailly*** (***D-B***) présente les résultats dus à l'enrichissement des synonymes contextuels conjoints des dictionnaires *Dubois* et *Bailly*. Aucun enrichissement aveugle n'est effectué ;
- le champ ***Dubois - EuroWordNet - Memodata*** (***D-EWN-M***) présente les résultats dus à l'enrichissement des synonymes contextuels conjoints des dictionnaires *Dubois*, *EuroWordNet* et *Memodata*. Aucun enrichissement aveugle n'est effectué ;
- la ligne ***Dubois - Dérivés*** (***D-Dér***) indique l'apport de la dérivation morphologique couplée à l'enrichissement contextuel des synonymes du dictionnaire *Dubois*. Aucun enrichissement aveugle n'est effectué ;
- le champ ***Dubois - Bailly - EuroWordNet - Memodata - Dérivés*** (***D-B-EWN-M-Dér***) affiche les résultats de l'ensemble des enrichissements possibles, exception faite de l'enrichissement synonymique aveugle des unités lexicales qui n'ont pas été désambiguïsées ;
- enfin, la ligne ***Tous types*** rassemble tous les modes d'enrichissement issus de méthodes linguistiques. Les enrichissements de la ligne précédente y sont effectués, auxquels est adjoint l'enrichissement aveugle des unités lexicales non désambiguïsées.

L'examen de ces résultats confirme l'apport important des traitements linguistiques, tant au niveau du score qu'à celui du nombre de réponses apportées. On voit principalement l'intérêt de la synonymie ciblée (*D*) par rapport à son équivalent aveugle (*Synonymes*). L'intérêt de la syntaxe seule n'est pas évident. Il pourrait être montré s'il ressort que cet apport n'est pas uniquement issu de la résolution de la coréférence (*cf.* table 7.3 page 212). Toutefois, l'analyse syntaxique n'est utilisée ici que comme support à d'autres enrichissements. Par exemple, il est évident que l'enrichissement synonymique contextuel est plus efficace si une analyse syntaxique est effectuée (*Synonymes*) que lorsque ce type d'enrichissement est effectué sur les simples mots-clefs extraits d'une requête (*Syn. aveugles*). Trois autres points intéressants ressortent de ces résultats.

Tout d'abord, nous constatons le faible apport de l'enrichissement issu de la dérivation morphologique (*D-Dér*). L'analyse des questions à laquelle nous avons procédé lors de la constitution de l'ensemble des requêtes à utiliser

Enrichissement	Score	Sans réponse
Plancher	0,295	139
Syn. aveugles	0,303	137
Syntaxe	0,360	127
Synonymes	0,369	125
D	0,388	121
D-Syn	0,388	121
D-B	0,393	120
D-EWN-M	0,413	116
D-Dér	0,393	120
D-B-EWN-M-Dér	0,428	113
Tous types	0,428	113

TAB. 7.2 – Résultats des planchers et de la méthode globale selon le protocole de question-réponse.

dans le protocole d'évaluation nous a permis de déceler une particularité qui est peut-être à l'origine de cet insuccès. En effet, les verbes dans les énoncés des requêtes correspondent fréquemment à des dérivation de noms ou d'adjectifs, présents dans plusieurs cas à l'intérieur de leurs réponses. Nous avons suffisamment signalé la richesse de l'information liée aux verbes par rapport à celle rattachée aux autres catégorie grammaticale dans le dictionnaire *Dubois*, à l'origine de l'enrichissement dérivationnel. Or il est fréquent que des dérivations mentionnées au départ d'entrées verbales ne rencontrent pas leur pendant à partir d'une autre catégorie grammaticale vers un dérivé verbal. Par exemple, alors qu'un lien de dérivation existe entre *protéger* et *protecteur* au départ du verbe, aucun lien équivalent n'est indiqué au départ de *protecteur* vers *protéger*. Du fait de cette carence, particulièrement sensible pour les noms et les adjectifs, de nombreux enrichissements ne peuvent s'effectuer. De ce fait, il arrive que des réponses ne soient pas sélectionnées lors de l'interrogation.

A contrario, nous remarquons la bonne tenue de l'information des dictionnaires sémantiques (*EuroWordNet* et *Memodata*), surtout par rapport à l'enrichissement apporté par le dictionnaire *Bailly*. Nous avions, il est vrai, souligné le peu d'intérêt de l'information de ce dictionnaire lorsque nous l'avons passé en revue (cf. section 3.4.2 page 119). Malgré cette faiblesse du dictionnaire *Bailly*, l'impact des synonymes contextuels sur la qualité des résultats est parfaitement évidente. Les courbes présentées dans la figure 7.1 page suivante permettent de confirmer de manière plus visuelle l'intérêt d'une analyse linguistique et l'apport essentiel d'une bonne synonymie contextuelle.

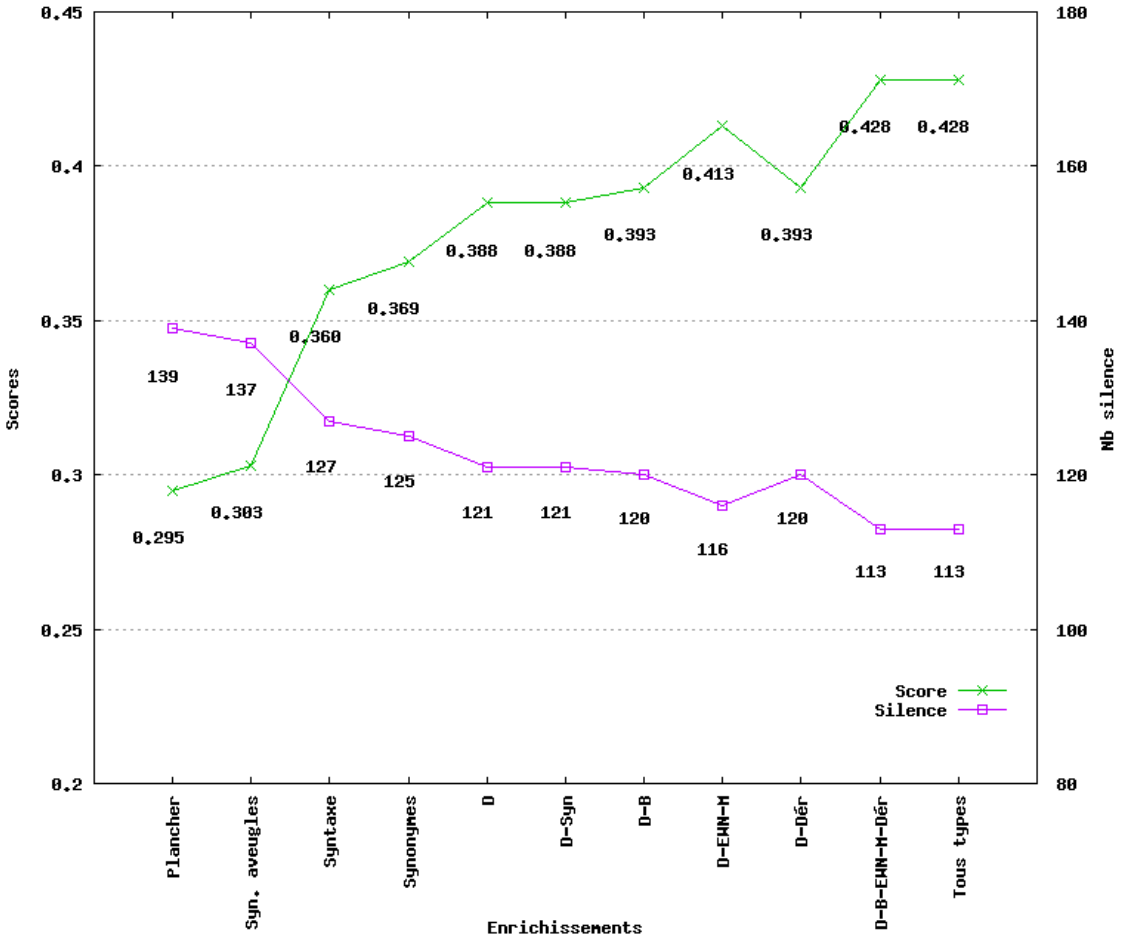

FIG. 7.1 – Performances du système global en question-réponse.

Enfin, l'addition de synonymes aveugles sur les lexèmes non désambiguïsés ne semble étrangement pas apporter la moindre amélioration. En effet, l'adjonction de ces synonymes aveugles à l'enrichissement synonymique du *Dubois*, comme à l'ensemble des autres enrichissements, n'améliore en rien les performances du système. Toutefois, nous n'avons pu effectuer cet enrichissement aveugle qu'à l'aide de l'information synonymique propre au *Dubois*, à l'exclusion des autres dictionnaires dans lesquels nous puisons l'information synonymique. Cet enrichissement peut être de piètre qualité, surtout dans les catégories non verbales. Cela peut justifier la stagnation constatée.

Les conclusions sur les différents types d'enrichissement que nous avons tirées des précédents résultats doivent à présent être vérifiées. Validées ou invalidées sur la méthode globale, ces techniques doivent à présent être soumises à des variations de paramètres pour en constater les fluctuations. Pour

ce faire, nous avons d'abord neutralisé le module de résolution de coréférence, ce qui va nous permettre de vérifier également l'importance de cet outil dans le processus de structuration de l'information.

	Sans coréférence		Avec coréférence	
Enrichissement	**Score**	**Sans réponse**	**Score**	**Sans réponse**
Plancher	0,295	139	—	—
Syn. aveugles	0,303	137	—	—
Syntaxe	0,275	144	0,360	127
Synonymes	0,284	142	0,369	125
D	0,304	138	0,388	121
D-Syn	0,304	138	0,388	121
D-B	0,303	138	0,393	120
D-EWN-M	0,324	134	0,413	116
D-Dér	0,304	138	0,393	120
D-B-EWN-M-Dér	0,328	133	0,428	113
Tous types	0,328	133	0,428	113

TAB. 7.3 – Résultats comparés de la méthode globale avec et sans coréférence.

Le tableau 7.3 présente les différents résultats issus des divers enrichissements, mais privés de la résolution de la coréférence. En comparant les résultats de l'interrogation de la méthode globale avec et sans coréférence, on peut constater l'ampleur de l'impact de cette coréférence sur l'efficacité du système : pour chacun des enrichissements, la perte de qualité est de plus de 0,075 point (soit près de 25 %) et le nombre de questions qui n'obtiennent pas de réponse croît d'approximativement 20 unités (presque 20 %). L'intérêt de l'utilisation d'un module de coréférence, même aussi rudimentaire que la technique que nous avons utilisée, est clairement établi. Nous traiterons ultérieurement de l'extension de cette coréférence aux adjectifs possessifs et aux autres pronoms.

Nous déduisons de ce tableau que l'apport d'une simple analyse morpho-syntaxique ne bénéficie pas à la tâche de question-réponse. Toutefois, le nombre de questions qui ont obtenu une réponse uniquement par la syntaxe est de 56, ce qui donne un score maximal de 0,280. Le score de 0,275 obtenu est donc presque parfait, proportionnellement plus élevé que celui du plancher, pourtant déjà excellent, de 0,295 pour un maximum de 0,305. Nous verrons dans les mesures traditionnelles [13] l'impact de la syntaxe sur la précision (*cf.* section 7.3.2 page 217).

[13] Nous appelons mesures traditionnelles, ou résultats traditionnels, les chiffres obtenus par le calcul de la précision, du rappel et de la F-mesure, qui sont utilisés dans toutes les tâches de la gestion de l'information depuis des décennies, excepté celle de question-réponse.

L'examen des différentes valeurs présentées dans ce tableau ne montre pas de variation particulière d'un type d'enrichissement par rapport aux autres dans une proportion réellement différente de celles de la méthode globale. Ce *statu quo* nous conforte donc dans nos observations concernant les différents enrichissements : importance de l'enrichissement synonymique contextuel, faiblesse relative de la dérivation morphologique pour ce type de questions et peu d'intérêt de l'enrichissement aveugle des lexèmes non désambiguïsés.

Toutefois, il est possible que l'exploitation ou non de la coréférence des pronoms sujets ne soit pas de nature à révéler les variations dans les possibilités des différents types d'enrichissements. Il est donc important d'étudier le comportement de ces enrichissements dans des situations différentes afin de pouvoir juger de leur apport réel. À partir de textes tantôt soumis à la résolution de la coréférence et tantôt non, nous allons faire varier les paramètres liés à la mise en correspondance des questions et réponses (utilisation ou non de l'unité lexicale *focus*, utilisation et variation du seuil de correspondance des dépendances contenues dans les réponses candidates et dans les questions).

Le système appliqué sur les textes avec résolution de la coréférence présente les meilleurs résultats si les réponses candidates qui ne présentent pas de correspondance de dépendance avec la question ne sont pas rejetées. Ils sont également les meilleurs dans le cas où la coréférence n'est pas résolue. Ces résultats apparaissent dans la table 7.4.

	Avec coréférence		Sans coréférence	
Enrichissement	**Score**	**Sans réponse**	**Score**	**Sans réponse**
Syntaxe	0,365	126	0,280	143
Synonymes	0,379	123	0,289	141
D	0,398	119	0,309	137
D-Syn	0,398	119	0,309	137
D-B	0,403	118	0,308	137
D-EWN-M	0,423	114	0,329	133
D-Dér	0,403	118	0,309	137
D-B-EWN-M-Dér	0,438	111	0,333	132
Tous types	0,438	111	0,333	132

TAB. 7.4 – Interrogation sans rejet des réponses sans correspondance syntaxique avec la question : résultats avec et sans coréférence.

Nous observons que dans ce cas également, les constatations que nous avons faites sur l'apport respectif des différents enrichissements restent d'actualité, que cet enrichissement soit synonymique contextuel ou aveugle, dérivationnel ou qu'il provienne de la résolution de la coréférence. Les résultats

présentés par les textes avec coréférence en particulier comptent parmi les meilleurs que nous avons obtenus. Près de la moitié des questions obtiennent une réponse dans les cinq premières propositions, et les résultats présentés quel que soit l'enrichissement sont supérieurs à tous les autres.

Nous relevons particulièrement la réaction inattendue du système à la demande de correspondance exacte des dépendances syntaxiques. En effet, les résultats sont meilleurs lorsque la contrainte syntaxique diminue. Pour évaluer l'impact des contraintes syntaxiques sur les résultats, nous présentons dans le tableau 7.5 les résultats de la méthode globale, mais nous faisons varier le seuil du rejet des réponses qui ne concordent pas totalement avec la question. Les résultats pour un seuil de 10 %, identiques à ceux du seuil à 0 %, n'ont pas été présentés, de même que ceux de 30 %, identiques aux résultats du seuil à 20 %.

	Seuil = 0 %		Seuil = 20 %		Seuil = 40 %	
Enrichissement	**Score**	**Sans rép.**	**Score**	**Sans rép.**	**Score**	**Sans rép.**
Syntaxe	0,360	127	0,355	128	0,335	132
Synonymes	0,369	125	0,364	126	0,344	130
D	0,388	121	0,378	123	0,358	127
D-Syn	0,388	121	0,378	123	0,358	127
D-B	0,393	120	0,383	122	0,363	126
D-EWN-M	0,413	116	0,403	118	0,378	123
D-Dér	0,393	120	0,383	122	0,363	126
D-B-EWN-M-Dér	0,428	113	0,418	115	0,388	121
Tous types	0,428	113	0,418	115	0,388	121

TAB. 7.5 – Résultats de la méthode globale avec variation du taux minimal de concordance entre question et réponse.

Ces résultats montrent que l'impact de l'élévation du seuil de concordance des dépendances sur les performances du système est plutôt négatif, même si cette variation n'a pas d'impact entre 0 % et 20 %, ni entre 20 % et 40 %. La diminution de qualité des réponses lors de l'augmentation des contraintes sur la syntaxe est surprenante, mais elle peut être justifiée. En effet, le renforcement d'une contrainte implique généralement la diminution des propositions de réponses. À partir de ce fait, la détérioration des résultats s'explique par la conjonction de deux facteurs : une réponse dont les dépendances ne correspondent pas à celles de la question est normalement placée très bas dans la liste des propositions, et ne donnera lieu à un résultat que si aucune réponse correcte n'est placée plus haut, mais ne fera pas évoluer favorablement le score si elle est rejetée ; d'autre part, une dépendance syntaxique présente dans une question propose un schéma très strict qui peut ne s'adapter que partiellement à la réalité du document qui y correspond.

De fait, la rigueur de ce schéma peut poser un problème dans le cadre d'énoncés où une dépendance sémantique de même type unit les deux mêmes entités, mais que ce lien n'existe au niveau syntaxique que de manière indirecte. Par exemple dans l'énoncé

Julius Caesar était le neveu de l'empereur Tibère.

extrait de l'*Encyclopédie Hachette Multimédia*, aucune dépendance ne permet de relier directement *neveu* et *Tibère*. En effet, le rattachement prépositionnel unit *neveu* et *empereur* (`NARG[INDIR](neveu,de,empereur)`), tandis que le mot *empereur* est lui-même relié à *Tibère* en tant qu'apposition du nom (`NN(empereur,Tibère)`). Une solution pour régler ce type de problème serait de considérer des dépendances indirectes permettant de définir un même lien sémantique comme correspondant à celles de la requête, même si cette correspondance est effectuée à un moindre degré.

Un autre cas où la rigueur du schéma syntaxique peut poser un problème à la mise en correspondance de la question et de la réponse vient de lacunes dans l'analyse syntaxique fournie par la grammaire. De fait, si nous prenons une phrase comme

Marc Antoine fut l'ami et le second de César.

extraite de l'*Encyclopédie Hachette Multimédia*, l'analyse syntaxique de *XIP* ne comportera qu'un seul rattachement prépositionnel impliquant *César* : `NMOD[INDIR](second,de,César)`. Une dépendance prépositionnelle impliquant *ami* et *César* dans une question ne pourrait donc être mise en correspondance avec la phrase présentée.

Dans ces deux cas, l'information lexicale est présente ainsi que l'information syntaxique. Le problème vient de ce que ces deux informations ne sont pas forcément compatibles avec celles des questions proposées, dont la syntaxe est souvent plus simple, et ne comporte donc généralement ni lacune ni de relation indirecte.

L'absence de l'unité lexicale correspondant au *focus* est le dernier paramètre sur lequel nous pouvons agir pour améliorer certains aspects des résultats produits par le système. Le tableau 7.6 page suivante présente les différents résultats obtenus en retirant le lexème porteur du *focus* des exigences de la question et en faisant varier les contraintes sur le seuil de concordance des dépendances des réponses et de la question : aucune exigence, suppression des concordances nulles, seuil de 10 % à 40 %.

De ces résultats, nous pouvons déduire que les réponses à concordance basse (sous 20 %) n'influencent que faiblement la qualité des résultats. Les courbes de résultats (*cf.* figure 7.2 page suivante) permettent de mieux percevoir l'impact des variations du seuil sur les performances. Des intervalles relativement importants entre les taux qui modifient réellement les résultats proviennent probablement de ce que la plupart des questions sont courtes, et

Seuil	Score	Sans réponse
Sans seuil	0,504	97
0 %	0,489	100
10 %	0,484	101
20 %	0,448	109
30 %	0,433	112
40 %	0,403	118

TAB. 7.6 – Résultats de la méthode sans *focus* avec variation du taux minimal de concordance entre question et réponse.

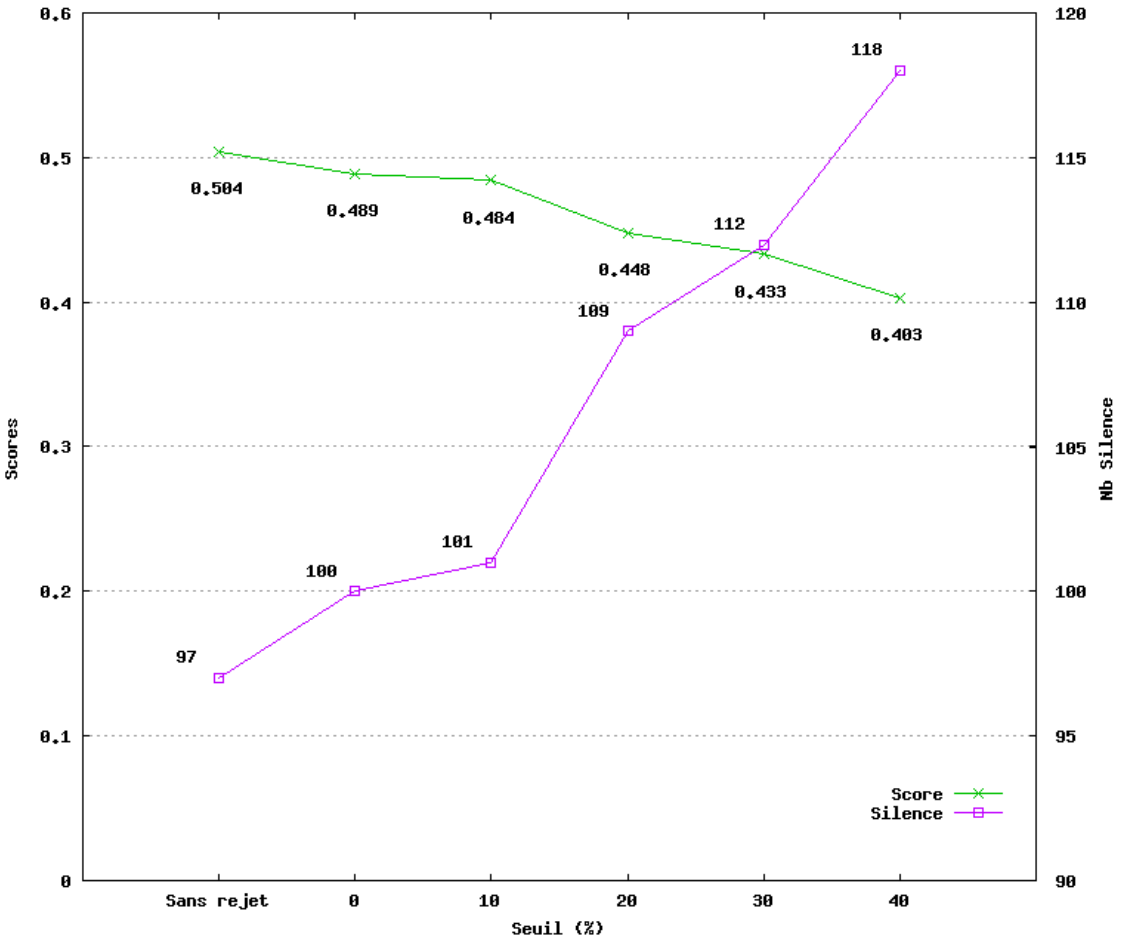

FIG. 7.2 – Impact de la variation du seuil de rejet sur les performances du système en question-réponse.

comportent donc peu de dépendances. Les variations de taux de correspondances sont peu nombreuses entre 0 % et 20 % et entre 20 % et 40 % lorsque peu de dépendances sont mises en jeu. Il est donc plus courant d'obtenir des taux de correspondance nuls ou échelonnés de 20 % en 20 %.

Il est cependant paradoxal de constater que les scores sont supérieurs lorsque les contraintes sont diminuées, tant par la suppression du *focus* [14] que par la diminution des contraintes syntaxiques sous la forme d'un seuil bas ou inexistant. Toutefois, l'augmentation du nombre des propositions de réponses alliée à une classification des réponses où les contraintes syntaxiques ont la prépondérance peut expliquer ce comportement inattendu.

Il serait intéressant de pouvoir tester le même système en le modifiant de manière à ce que la ou les dépendances qui comportent le *focus* soient obligatoirement retrouvées dans les réponses candidates, surtout dans les cas où l'unité lexicale désignée par le *focus* n'est pas exigée.

De plus, pour l'ensemble des derniers résultats présentés, le détail des enrichissements nous permet de confirmer les conclusions que nous tirions à propos de leur intérêt respectif. En particulier, nous constatons le peu d'efficacité de la morphologie dérivationnelle et de l'enrichissement aveugle des lexèmes non désambiguïsés, tandis que l'importance de la coréférence est confirmée. Dans tous les cas, l'apport d'un enrichissement contextuel lié au sens est également un atout majeur. Par ailleurs, la classification des réponses liée à la syntaxe semble un moyen efficace de conserver de fait certaines contraintes syntaxiques sur les réponses proposées.

7.3.2 Les résultats traditionnels de la gestion de l'information

Les critères d'évaluation intéressants pour les autres disciplines liées à la gestion de l'information ne sont pas forcément identiques à ceux qui servent à évaluer la tâche de question-réponse. En particulier, toutes les réponses possibles y sont généralement prises en compte : rappel, ainsi que toutes les réponses fournies : précision (*cf.* section 1.2.1 page 26). Le corpus que nous traitons comporte 249 réponses correctes pour les 200 questions qui lui sont proposées. Malgré ces différences, les paramètres que nous pouvons faire varier sont identiques à ceux que nous avons manipulés pour évaluer les capacités de question-réponse du système, excepté la possibilité de travailler sur des fenêtres plus larges (paragraphe ou texte). Les conclusions que nous pouvons tirer sont souvent semblables. Nous nous attardons donc peu sur les phénomènes déjà constatés, tandis que nous insistons sur les particularités

[14] Encore que la présence de l'unité lexicale désignée par le *focus* ne se justifie pas toujours dans la réponse. En effet, dans des questions du type *Quelle est la couleur des yeux de César ?*, le *focus* qui correspond au mot *couleur* n'est pas présent dans la réponse.

propres à ce type de calcul.

La méthode globale permet d'isoler les résultats présentés dans le tableau 7.7. Nous constatons immédiatement que la précision atteinte par le plancher est excellente, mais que ce résultat est tempéré par un rappel bas. Le calcul de la F-mesure permet de ne pas s'y tromper. En effet, alors que cette F-mesure est honorable lorsque la précision est favorisée (β=0,5), elle s'effondre de près de 20 points dès lors qu'il y a équivalence entre la précision et le rappel (β=1), et chute encore si le rappel est privilégié (β=2). La méthode globale n'atteint pas le même niveau de précision, bien que cette précision soit de plus de 38 % et qu'elle permette d'obtenir une mesure d'ensemble supérieure dans tous les cas à la mesure du plancher, même lorsque la précision est favorisée.

Enrichissement	Précision	Rappel	F-m1	F-m2	F-m3
Plancher	83.75%	26.91%	58.88	40.73	31.13
Syn. aveugles	80.90%	28.92%	59.50	42.60	33.18
Tous types	79.83%	38.15%	65.52	51.63	42.60

TAB. 7.7 – Résultats traditionnels des méthodes de plancher et de la méthode globale.

Le tableau quantitatif 7.8 et les courbes d'évaluation de la figure 7.3 page suivante permettent de confirmer ces chiffres : le nombre de bonnes réponses apportées par notre méthode dépasse le plancher de plus de 41 %, et le nombre de questions qui obtiennent au moins une bonne réponse augmente de plus de 40 %. Il est vrai que notre méthode double le nombre de réponses fausses avec 24 cas là où la mesure de plancher n'en a que 13. Cette différence justifie en partie la différence de précision entre le plancher et notre méthode globale. Les réponses rejetées (***Rejet***) sont les candidates dont le taux de concordance des dépendances avec celles de la question est inférieur au seuil prescrit (ici 0 %). Parmi ces réponses rejetées, nous indiquons le nombre de propositions correctes (***Rej. exact***).

Enrichissement	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Plancher	67	59	13	136	—	—
Syn. aveugles	72	61	17	133	—	—
Tous types	95	83	24	109	13	3

TAB. 7.8 – Résultats quantitatifs traditionnels de la méthode de plancher et de la méthode globale.

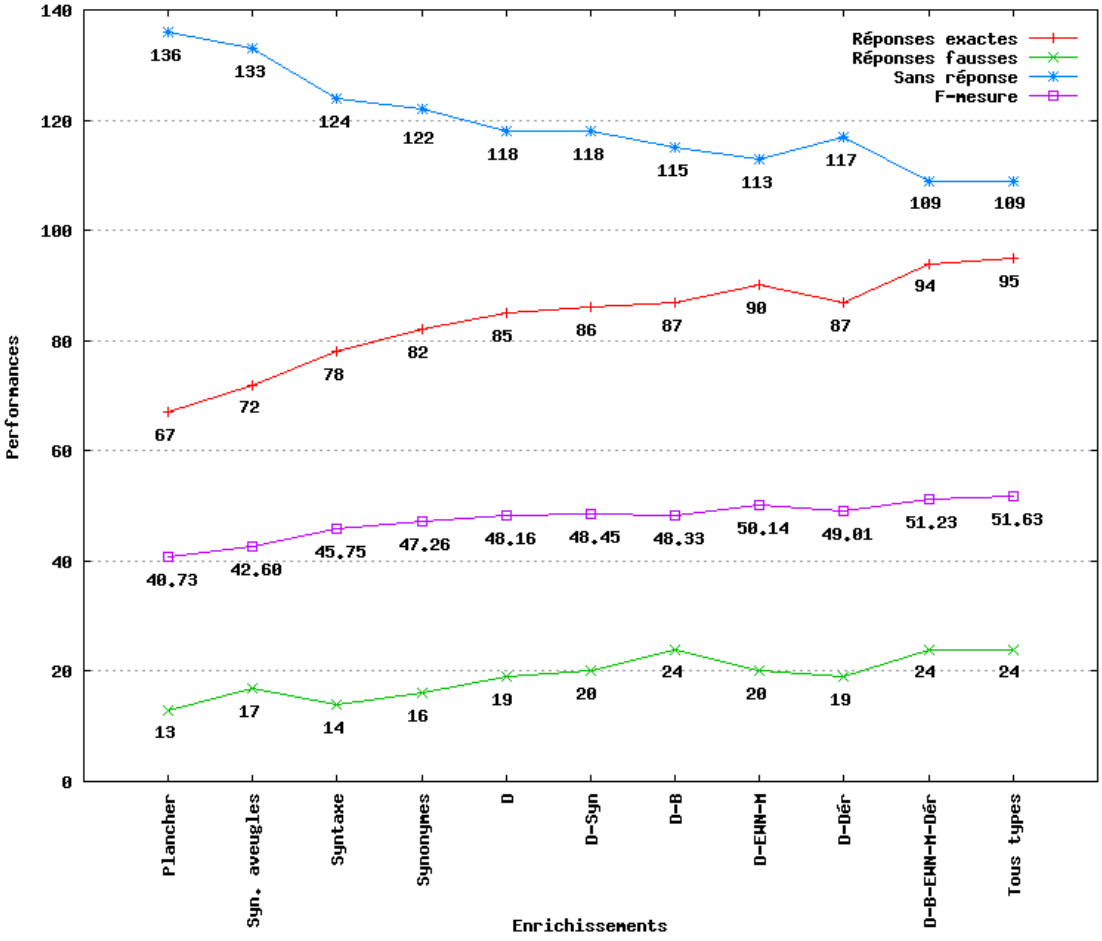

FIG. 7.3 – Courbes des performances quantitatives de la méthode globale.

Les caractéristiques du corpus d'évaluation, choisi délibérément parmi des articles encyclopédiques traitant de personnages pour les raisons déjà évoquées, sont à l'origine de la qualité des réponses de la méthode de plancher qui est particulièrement élevée, et peut-être, dans une mesure que nous espérons moindre, des autres résultats également. Ce corpus contient en effet une grande quantité de noms propres, ce qui est fréquent dans une encyclopédie, plus encore si les documents portent sur des personnes. Comme une grande majorité des questions contient au moins un nom propre (93 %), la tâche de recherche d'information par mots-clefs s'en trouve facilitée [15].

[15] Étant donné que les noms propres qui ne sont pas recensés dans le lexique sont fréquemment à l'origine d'erreurs lors de l'analyse syntaxique des textes, notre approche ne bénéficie pas systématiquement du même avantage. Or les noms de personnages historiques romains ne sont qu'occasionnellement catalogués dans ce lexique, mais peuvent fréquemment être considérés comme des mots appartenant au lexique commun (*Pompée*,

Les résultats obtenus dans cette évaluation ne seront donc pas forcément représentatifs de la qualité des méthodes sur d'autres types de textes ou de sujets. Toutefois, les processus mis en œuvre dans notre méthode peuvent être testés les uns par rapport aux autres et nous avons donc la possibilité de déterminer leurs intérêts respectifs. Par ailleurs, notre méthode est destinée à une utilisation généraliste, et ne doit donc pas récuser un type de texte particulier, qu'il lui soit favorable ou défavorable, comme c'est ici le cas.

La présentation des résultats de la méthode globale détaillés en fonction du processus d'enrichissement dans le tableau 7.9 permet de confirmer certaines conclusions que nous avons tirées lors de l'analyse des données obtenues dans la perspective de la tâche de question-réponse.

Tout d'abord, nous constatons ici l'apport de l'analyse morpho-syntaxique en comparaison avec les résultats du plancher. Le bénéfice est manifeste du point de vue de la précision, comme nous l'avions pressenti sans pouvoir le démontrer dans les chiffres issus de la tâche de question-réponse. L'augmentation du rappel est due à l'utilisation de la coréférence, mais ce type d'enrichissement ne privilégie pas la précision. Le tableau quantitatif (*cf.* table 7.10 page suivante) des mêmes résultats confirme la qualité de l'information identifiée au travers de cette seule analyse, qui voit augmenter significativement le nombre de ses bonnes réponses grâce à la résolution anaphorique sans accroître en proportion le nombre des erreurs. L'analyse des résultats sans application de la coréférence (*cf.* table 7.11 page 222) ne se démarque pas de cette observation.

Enrichissement	Précision	Rappel	F-m1	F-m2	F-m3
Plancher	83.75%	26.91%	58.88	40.73	31.13
Syn. aveugles	80.90%	28.92%	59.50	42.60	33.18
Syntaxe	84.78%	31.33%	63.21	45.75	35.85
Synonymes	83.67%	32.93%	63.96	47.26	37.48
D	81.73%	34.14%	63.91	48.16	38.64
D-Syn	81.13%	34.54%	63.89	48.45	39.02
D-B	78.38%	34.94%	62.77	48.33	39.30
D-EWN-M	81.82%	36.14%	65.31	50.14	40.69
D-Dér	82.08%	34.94%	64.64	49.01	39.47
D-B-EWN-M-Dér	79.66%	37.75%	65.19	51.23	42.19
Tous types	79.83%	38.15%	65.52	51.63	42.60

TAB. 7.9 – Résultats traditionnels détaillés des planchers et de la méthode globale.

Galère, *César*, *Auguste*, etc.).

La faiblesse de deux enrichissements est également confirmée par ces résultats, aussi bien par les mesures de précision et de rappel que par les mesures quantifiées. Il s'agit des enrichissements apportés par la synonymie aveugle et par la morphologie dérivationnelle. Les apports de la synonymie aux lexèmes qui n'ont pas été désambiguïsés sont anecdotiques : en précision comme en rappel, elle est de 1 % et l'amélioration quantitative n'est pas plus convaincante. La dérivation morphologique est tout aussi décevante, qu'elle soit appliquée après la seule désambiguïsation ou après l'adjonction de tous les autres enrichissements. Ces enrichissements ne sont donc pas très significatifs dans notre méthode.

Enrichissement	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Plancher	67	59	13	136	0	0
Syn. aveugles	72	61	17	133	0	0
Syntaxe	78	70	14	124	5	1
Synonymes	82	72	16	122	9	2
D	85	76	19	118	11	3
D-Syn	86	76	20	118	11	3
D-B	87	77	24	115	12	3
D-EWN-M	90	81	20	113	11	3
D-Dér	87	77	19	117	11	3
D-B-EWN-M-Dér	94	83	24	109	12	3
Tous types	95	83	24	109	13	3

TAB. 7.10 – Résultats traditionnels quantitatifs détaillés de la plancher et de la méthode globale.

Par contre, nous constatons les bons résultats obtenus par l'enrichissement synonymique contextuel en général, avec une meilleure efficacité pour les dictionnaires sémantiques (*EuroWordNet* et *Memodata*) que pour le dictionnaire synonymique (*Bailly*). Le tableau des mesures quantitatives permet d'envisager que les bonnes réponses dont l'obtention est permise par les synonymes issus du *Bailly* (deux de plus que l'enrichissement du seul *Dubois*), par les dictionnaires sémantiques (cinq de plus que l'enrichissement du *Dubois*) et par la dérivation morphologique (deux de plus que l'enrichissement du *Dubois*) ne se recouvrent pas. Nous constatons aussi que, si la dérivation morphologique n'offre somme toute qu'un apport modeste au système, sa contribution n'en est pas moins d'excellente qualité, car aucune erreur n'est générée par cet enrichissement. Chacun de ces enrichissements a donc sa raison d'être utilisé dans notre méthodologie.

Les chiffres des tableaux 7.11 page suivante et 7.12 page suivante présentent les résultats de l'interrogation de la base en exploitant les mêmes

processus – excepté la technique de résolution de coréférence des pronoms sujets – et les mêmes paramètres, à savoir la présence de l'unité lexicale désignée par le *focus* dans la réponse ainsi que l'élimination des réponses qui ne présentent aucune dépendance concordante avec la requête. Ces résultats permettent de confirmer l'importance de l'analyse syntaxique, qui conserve un avantage sur les résultats du plancher grâce à une précision meilleure tandis que la perte en rappel reste relativement faible.

L'analyse quantitative surtout permet de constater l'amélioration apportée par la syntaxe (diminution des réponses fausses de 31 % pour une diminution de 12 % des réponses fournies). L'examen des autres résultats permet de légitimer nos appréciations des autres enrichissements les uns par rapport aux autres et de constater que les réponses apportées par l'enrichissement synonymique issu des différents types de dictionnaires ne se recoupent pas.

	Sans coréférence					Avec coréférence				
Enrichissement	**Précision**	**Rappel**	**F-m1**	**F-m2**	**F-m3**	**Précision**	**Rappel**	**F-m1**	**F-m2**	**F-m3**
Syntaxe	86.76%	23.69%	56.62	37.22	27.73	84.78%	31.33%	63.21	45.75	35.85
Synonymes	85.14%	25.30%	57.80	39.01	29.44	83.67%	32.93%	63.96	47.26	37.48
D	82.28%	26.10%	57.52	39.63	30.23	81.73%	34.14%	63.91	48.16	38.64
D-Syn	81.48%	26.51%	57.59	40.00	30.64	81.13%	34.54%	63.89	48.45	39.02
D-B	78.57%	26.51%	56.41	39.64	30.56	78.38%	34.94%	62.77	48.33	39.30
D-EWN-M	81.18%	27.71%	58.57	41.32	31.91	81.82%	36.14%	65.31	50.14	40.69
D-Dér	82.28%	26.10%	57.52	39.63	30.23	82.08%	34.94%	64.64	49.01	39.47
D-B-EWN-M-Dér	76.92%	28.11%	57.10	41.18	32.20	79.66%	37.75%	65.19	51.23	42.19
Tous types	78.02%	28.51%	57.91	41.76	32.66	79.83%	38.15%	65.52	51.63	42.60

TAB. 7.11 – Résultats traditionnels comparés de l'utilisation d'un seuil de rejet des réponses sans concordance syntaxique avec la question nul ou minimal.

	Sans coréférence						Avec coréférence					
Enrichissement	**Exact**	**1 ex.**	**Faux**	**Sans**	**Rejet**	**R. ex.**	**Exact**	**1 ex.**	**Faux**	**Sans**	**Rejet**	**R. ex.**
Syntaxe	59	53	9	142	5	1	78	70	14	124	5	1
Synonymes	63	55	11	139	7	2	82	72	16	122	9	2
D	65	58	14	135	9	3	85	76	19	118	11	3
D-Syn	66	58	15	134	9	3	86	76	20	118	11	3
D-B	66	59	18	133	10	3	87	77	24	115	12	3
D-EWN-M	69	62	16	130	9	3	90	81	20	113	11	3
D-Dér	65	58	14	135	9	3	87	77	19	117	11	3
D-B-EWN-M-Dér	70	63	21	127	10	3	94	83	24	109	12	3
Tous types	71	63	20	128	10	3	95	83	24	109	13	3

TAB. 7.12 – Résultats traditionnels comparés de l'utilisation ou non de la résolution de la coréférence dans la méthode globale.

Nous pouvons à présent manipuler le seuil de concordance des dépendances des questions et des réponses, à partir duquel une phrase peut être acceptée comme réponse à la question posée. Le seuil testé précédemment était de 0 %, c'est-à-dire que la réponse était supprimée dès qu'aucune dépendance ne correspondait parfaitement entre la question et la réponse. Les tableaux 7.13 et 7.14 page suivante comparent les résultats obtenus lors de l'interrogation avec rejet des réponses sans concordance de dépendance avec la question, et ceux obtenus sans rejet. Contrairement au cas de figure obtenu lors de l'évaluation en question-réponse, les résultats obtenus en supprimant le seuil d'élimination de réponses non concordantes ne produisent pas un grand changement dans les résultats.

	Sans seuil					Avec seuil à 0 %				
Enrichissement	**Précision**	**Rappel**	**F-m1**	**F-m2**	**F-m3**	**Précision**	**Rappel**	**F-m1**	**F-m2**	**F-m3**
Syntaxe	83.51%	32.53%	63.58	46.82	37.05	84.78%	31.33%	63.21	45.75	35.85
Synonymes	81.31%	34.94%	64.25	48.88	39.44	83.67%	32.93%	63.96	47.26	37.48
D	78.26%	36.14%	63.47	49.45	40.50	81.73%	34.14%	63.91	48.16	38.64
D-Syn	77.78%	36.55%	63.46	49.73	40.88	81.13%	34.54%	63.89	48.45	39.02
D-B	74.80%	36.95%	62.08	49.46	41.11	78.38%	34.94%	62.77	48.33	39.30
D-EWN-M	78.51%	38.15%	64.80	51.35	42.52	81.82%	36.14%	65.31	50.14	40.69
D-Dér	78.63%	36.95%	64.16	50.27	41.33	82.08%	34.94%	64.64	49.01	39.47
D-B-EWN-M-Dér	76.15%	39.76%	64.37	52.24	43.96	79.66%	37.75%	65.19	51.23	42.19
Tous types	75.76%	40.16%	64.35	52.49	44.33	79.83%	38.15%	65.52	51.63	42.60

TAB. 7.13 – Résultats traditionnels comparés de l'utilisation ou non de la résolution de la coréférence dans la méthode globale.

Comme on pouvait le prévoir en diminuant les contraintes, la précision y perd un peu et le rappel y gagne, mais le calcul de ces deux mesures confondues (F-mesure avec β=1) donne globalement un résultat semblable. Cette différence avec les résultats de question-réponse vient de l'élimination naturelle d'un grand nombre de réponses à faible taux de concordance du fait de la classification des réponses les plus pertinentes de l'élimination des réponses les moins susceptibles d'être exactes dans la tâche de question-réponse. L'analyse des résultats quantifiés confirme ce raisonnement : les réponses correctes sont un peu plus nombreuses et les erreurs également.

Les variations du seuil de concordance en dessous duquel les réponses doivent être supprimées peuvent fluctuer. Les résultats avec un seuil à 0 % et à 10 % sont identiques, les questions étant généralement trop brèves pour contenir plus de dix dépendances. Les variations de résultats commencent à 20 %. Le tableau 7.15 page suivante présente les résultats pour un enrichissement qui comprend l'ensemble des méthodes décrites avec des seuils de 0 %, 10 %, 20 %, 30 % et 40 %, attendu que l'examen des différentes

Enrichissement	Sans seuil				Avec seuil à 0%					
	Exact	1 exact	Faux	Sans	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Syntaxe	81	72	16	122	78	70	14	124	5	1
Synonymes	87	75	20	119	82	72	16	122	9	2
D	90	79	25	115	85	76	19	118	11	3
D-Syn	91	79	26	115	86	76	20	118	11	3
D-B	92	80	31	112	87	77	24	115	12	3
D-EWN-M	95	84	26	110	90	81	20	113	11	3
D-Dér	92	80	25	114	87	77	19	117	11	3
D-B-EWN-M-Dér	99	86	31	106	94	83	24	109	12	3
Tous types	100	86	32	106	95	83	24	109	13	3

TAB. 7.14 – Résultats traditionnels quantitatifs comparés de l'utilisation d'un seuil de rejet des réponses sans concordance syntaxique avec la question nul ou minimal.

méthodes d'enrichissement n'apporte rien à nos constatations antérieures [16]. Le tableau 7.16 présente les résultats quantifiés correspondants.

Seuil	Précision	Rappel	F-m1	F-m2	F-m3
0 %	79.83%	38.15%	65.52	51.63	42.60
10 %	79.83%	38.15%	65.52	51.63	42.60
20 %	79.49%	37.35%	64.85	50.82	41.78
30 %	79.49%	37.35%	64.85	50.82	41.78
40 %	81.31%	34.94%	64.25	48.88	39.44

TAB. 7.15 – Résultats traditionnels des variations du seuil de concordance pour un enrichissement tous types confondus.

Seuil	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
0 %	95	83	24	109	13	3
10 %	95	83	24	109	13	3
20 %	93	81	24	111	15	3
30 %	93	81	24	111	15	3
40 %	86	76	20	117	24	3

TAB. 7.16 – Résultats traditionnels quantitatifs des variations du seuil de concordance pour un enrichissement tous types confondus.

[16] Nous avons ici favorisé la lisibilité, le confort et la facilité de comparaison entre les différents résultats. L'annexe C page 281 rassemble l'ensemble des tableaux de résultats. On pourra s'y reporter pour connaître le détail des apports de chacun des enrichissements.

L'évolution des résultats en fonction du renforcement des contraintes syntaxiques n'est pas très probant. En effet, l'évolution réelle des résultats demande un seuil très élevé, en dessous duquel ni la précision, ni le rappel ne bouge beaucoup. Il faut en effet un seuil de 40 % au moins pour voir une amélioration significative de la précision. Cette amélioration s'effectue au détriment du rappel, et cette perte du rappel est plus importante que l'apport en précision, comme le prouve la baisse de la F-mesure qui équilibre l'importance de la précision et du rappel. La hausse de ce seuil ne permet un gain de précision qu'au prix d'une perte plus importante du rappel. Le tableau quantitatif permet d'expliquer la faible variation des résultats avant un seuil de 40 % : les réponses rejetées sont relativement peu nombreuses avant cette limite de 40 %, et surtout leur nombre ne varie pas beaucoup. Mais dès le seuil de 40 %, la contrainte syntaxique est plus efficace et l'élimination de réponses, même de bonnes réponses[17], est plus importante.

Il nous reste maintenant à analyser les résultats produits par la variation du dernier paramètre que nous pouvons modifier, c'est-à-dire le *focus*. Dès lors que l'unité lexicale désignée par le *focus* n'est pas requise dans les réponses, les propositions sont beaucoup plus nombreuses du fait de l'élargissement des contraintes.

Les tableaux 7.17 page suivante et 7.18 page suivante comparent les résultats obtenus sans l'exigence de la présence du *focus* en faisant varier de nulle (aucun seuil de rejet) à faible (rejet des propositions avec 0 % de concordance) la contrainte du seuil de concordance des dépendances entre question et réponses. Ces résultats très peu contraints obtiennent un taux de rappel particulièrement élevé, au détriment de la précision. La F-mesure qui privilégie la précision (β=0.5) est très basse même si les autres scores sont honorables.

Les résultats quantitatifs indiquent par ailleurs que les réponses erronées sont très nombreuses. L'ajout d'une contrainte avec un seuil de concordance à 0 % n'améliore que faiblement les performances du système. Comme nous l'avons constaté précédemment, les contraintes basses sur les réponses peu concordantes, lorsque le système ne limite pas les propositions à un petit nombre, n'ont que peu d'impact sur les résultats. Un seuil de concordance de 10 % permet d'ailleurs d'obtenir des résultats identiques à ceux atteints lorsque le seuil est de 0 %.

Les résultats obtenus en appliquant des seuils plus élevés permettent d'obtenir un compromis acceptable entre les différentes mesures. Le graphique 7.4 page 227 qui présente les courbes quantitatives à chaque seuil testé permet d'évaluer visuellement l'impact des rejets sur les résultats. En

[17] Contrairement à l'évaluation de la tâche de question-réponse, les réponses les moins pertinentes d'un point de vue syntaxique n'ont pas été éliminées lors d'une sélection de cinq réponses.

Enrichissement	Sans seuil					Avec seuil à 0 %				
	Précision	Rappel	F-m1	F-m2	F-m3	Précision	Rappel	F-m1	F-m2	F-m3
Syntaxe	31.63%	42.17%	33.29	36.14	39.53	32.89%	40.16%	34.13	36.17	38.46
Synonymes	30.41%	44.58%	32.48	36.16	40.78	31.33%	41.77%	32.97	35.80	39.16
D	31.51%	46.18%	33.65	37.46	42.25	32.63%	43.37%	34.33	37.24	40.69
D-Syn	30.34%	46.18%	32.58	36.62	41.82	31.40%	43.37%	33.23	36.42	40.30
D-B	31.45%	46.99%	33.68	37.68	42.76	32.64%	44.18%	34.44	37.54	41.26
D-EWN-M	32.35%	48.19%	34.62	38.71	43.89	33.53%	45.38%	35.38	38.57	42.39
D-Dér	31.28%	46.99%	33.52	37.56	42.70	32.45%	44.18%	34.27	37.41	41.20
D-B-EWN-M-Dér	32.04%	49.80%	34.50	38.99	44.83	33.33%	46.99%	35.39	39.00	43.43
Tous types	31.00%	49.80%	33.53	38.21	44.41	32.23%	46.99%	34.39	38.24	43.05

TAB. 7.17 – Résultats traditionnels comparés de l'utilisation ou non d'un seuil de concordance sans la présence du lexème *focus*.

Enrichissement	Sans coréférence				Avec coréférence					
	Exact	1 exact	Faux	Sans	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Syntaxe	105	82	227	110	100	79	204	113	28	8
Synonymes	111	86	254	106	104	82	228	110	33	9
D	115	90	250	102	108	86	223	106	34	10
D-Syn	115	90	264	102	108	86	236	106	35	10
D-B	117	91	255	100	110	87	227	104	35	10
D-EWN-M	120	95	251	97	113	91	224	101	34	10
D-Dér	117	91	257	101	110	87	229	105	35	10
D-B-EWN-M-Dér	124	97	263	93	117	93	234	97	36	1
Tous types	124	97	276	93	117	93	246	97	37	10

TAB. 7.18 – Résultats traditionnels quantitatifs comparés de l'utilisation ou non d'un seuil de concordance sans la présence du lexème *focus*.

effet, un seuil de 20 % permet de retrouver un niveau de précision intéressant tout en conservant un rappel élevé étant donné que le nombre de propositions éliminées augmente très significativement. La F-mesure équilibrée (β=1) permet même de constater que c'est le meilleur compromis testé pour ce système. Les tableaux 7.19 page suivante et 7.20 page 228 contiennent les résultats comparés de l'interrogation de la base documentaire sans utilisation du *focus* dans la réponse et à des seuils de concordance inexistant et à 0 %, 10 %, 20 %, 30 % et 40 %. Seuls les résultats de tous les enrichissements confondus sont indiqués, le détail des différents enrichissements n'appelant pas de nouvelle constatation.

Ces tableaux confirment que le pic des résultats est bien atteint au seuil de 20 %. Au-delà, la précision augmente tandis que le rappel diminue, suite au renforcement des contraintes. En deçà de 20 %, les contraintes ne sont plus suffisantes pour garantir une précision acceptable, même au regard du

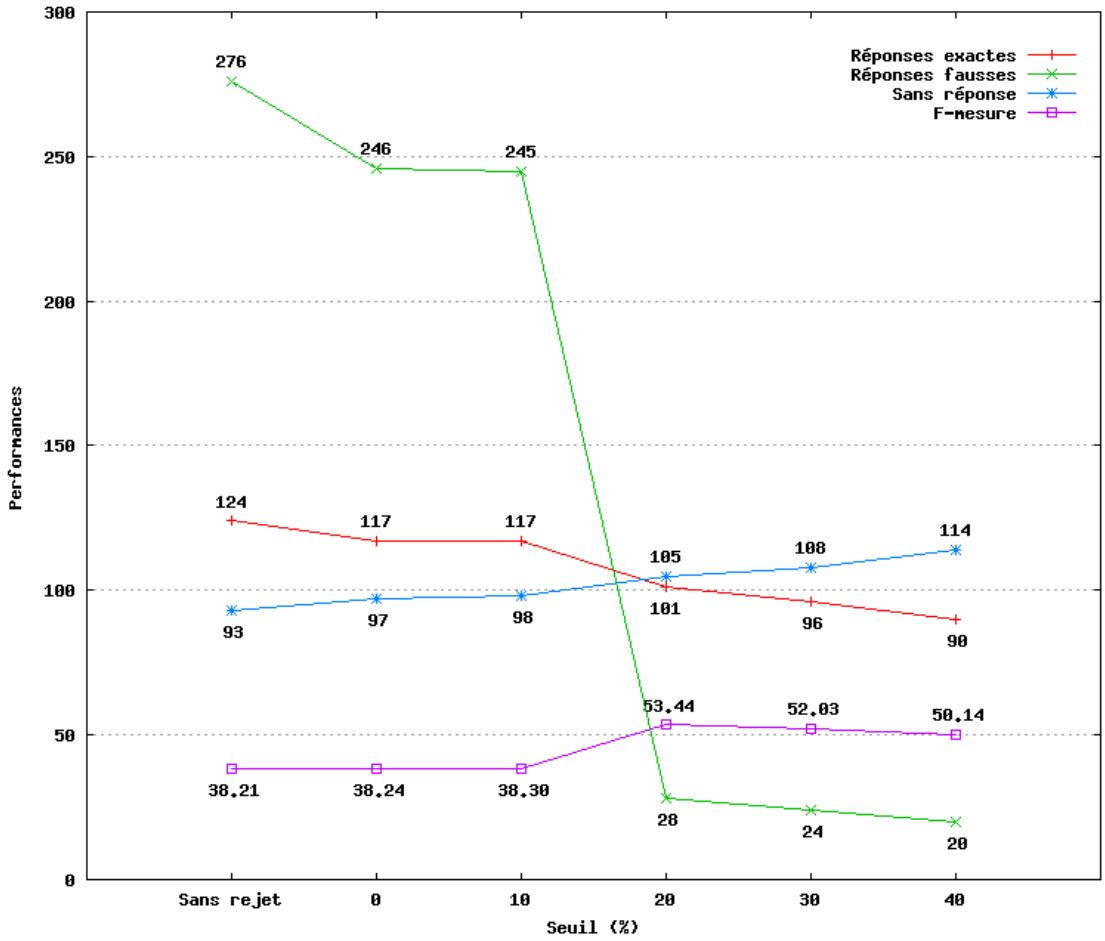

FIG. 7.4 – Influence du seuil de rejet sur les performances du système avec utilisation du *focus*.

Seuil	Précision	Rappel	F-m1	F-m2	F-m3
Sans seuil	31.00%	49.80%	33.53	38.21	44.41
0 %	32.23%	46.99%	34.39	38.24	43.05
10 %	32.32%	46.99%	34.47	38.30	43.08
20 %	78.29%	40.56%	66.01	53.44	44.89
30 %	80.00%	38.55%	65.84	52.03	43.01
40 %	81.82%	36.14%	65.31	50.14	40.69

TAB. 7.19 – Résultats traditionnels comparés des variations du seuil de concordance sans la présence du lexème *focus*.

Seuil	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans seuil	124	97	276	93	—	—
0 %	117	93	246	97	37	10
10 %	117	93	245	98	38	10
20 %	101	87	28	105	271	24
30 %	96	84	24	108	280	26
40 %	90	79	20	114	290	26

TAB. 7.20 – Résultats traditionnels quantitatifs comparés des variations du seuil de concordance sans la présence du lexème *focus*.

rappel important. Les indications quantitatives indiquent que le nombre des réponses rejetées ne devient significatif qu'à partir d'un seuil de 20 %, au regard du nombre d'erreurs proposées. La figure 7.5 page suivante montre bien l'influence bénéfique des contraintes de seuil sur la précision à partir de 20 % lorsque le *focus* n'est pas exploité.

7.3.3 Élargissement de la fenêtre en gestion de l'information

La fenêtre de réponse disponible dans le cadre d'autres applications de gestion de l'information que la tâche de question-réponse ne correspond pas forcément à une phrase. Notre méthode permet d'interroger également une base textuelle à d'autres niveaux, tant du document lui-même que du paragraphe, pour autant que cette notion soit définie au préalable.

Toutefois, nous ne disposons pas actuellement des mêmes fonctionnalités pour manipuler l'information dans ces fenêtres. En effet, nous ne pouvons pas traiter séparément les dépendances extraites lors de l'analyse de la base textuelle. De ce fait, il ne nous est pas possible d'établir un seuil de rejet des réponses dont le schéma syntaxique ne correspond pas suffisamment à la question. Par contre, nous pouvons gérer l'utilisation du lexème désigné par le *focus* ainsi que le module de résolution de coréférence des pronoms sujets.

Les tableaux 7.21 page 230 et 7.22 page 230 montrent les résultats comparés de l'interrogation de la base textuelle au niveau du paragraphe avec d'un côté l'utilisation du module de résolution de la coréférence et de l'autre la désactivation de ce même module. Par ailleurs, les tableaux 7.23 page 230 et 7.24 page 231 comparent les résultats obtenus pour l'ensemble des enrichissements confondus lorsque la présence du lexème *focus* est exigé dans la réponse et lorsqu'il ne l'est pas.

Les résultats obtenus montrent une nouvelle fois la qualité de l'enrichissement issu de la synonymie contextuelle, et principalement des dictionnaires

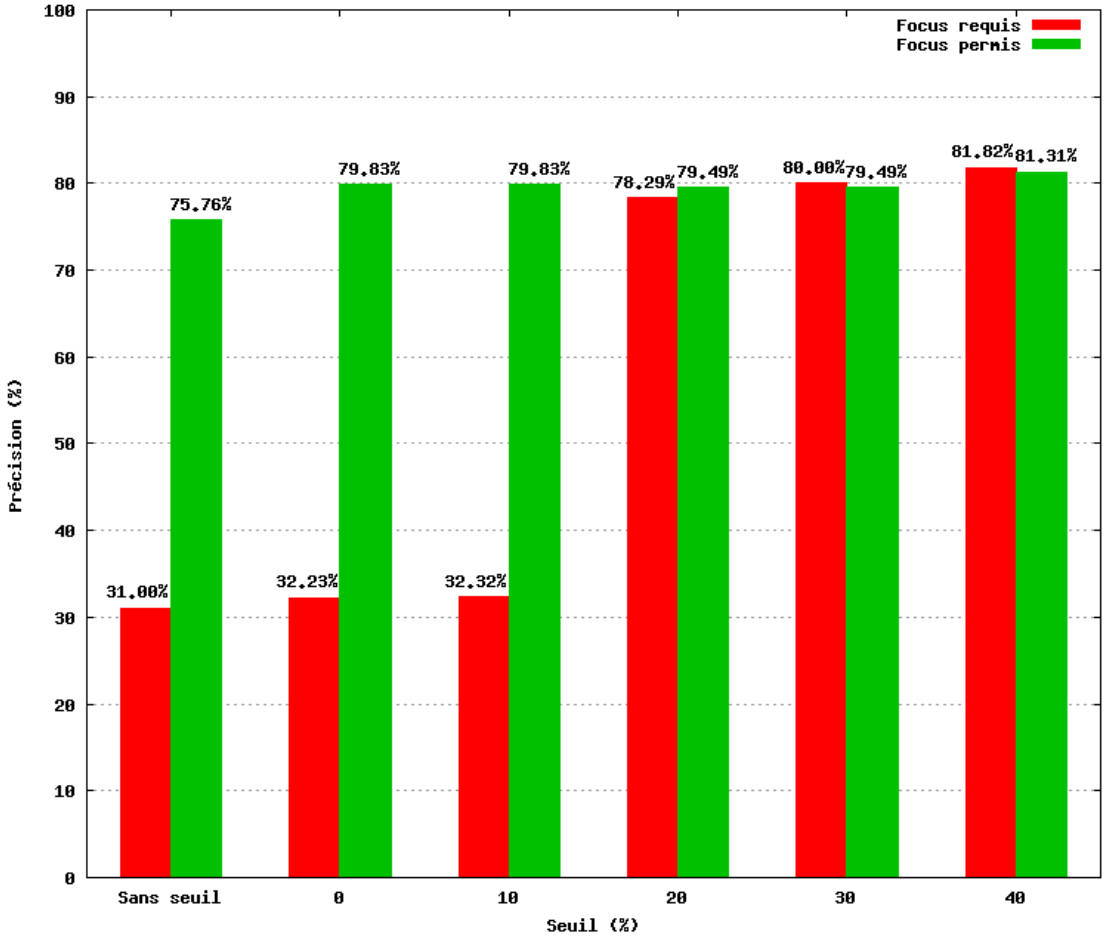

FIG. 7.5 – Évolution de la précision du système en fonction de la contrainte de seuil.

sémantiques. C'est en effet leur information qui permet d'atteindre le rappel le plus large tout en maintenant la précision à un degré élevé, quel que soit le paramètre modifié. Par ailleurs, l'enrichissement synonymique aveugle des lexèmes non désambiguïsés confirme la pauvreté de son apport, négatif dans tous les cas à ce niveau. Pour la première fois, l'enrichissement issu de la dérivation morphologique génère des erreurs en nombre plus important que des réponses correctes.

Par ailleurs, si l'importance du *focus* reste déterminante pour la précision, l'utilisation du module de résolution de coréférence est moins déterminante que lors de son utilisation dans des phrases. En effet, l'entité désignée par le pronom anaphorique est souvent présente dans le même paragraphe que ce pronom. L'utilisation de ce module est donc moins déterminante à ce niveau

	Sans coréférence					Avec coréférence				
Enrichissement	Précision	Rappel	F-m1	F-m2	F-m3	Précision	Rappel	F-m1	F-m2	F-m3
Plancher	83.02%	35.34%	65.38	49.58	39.93	—	—	—	—	—
Syn. aveugles	77.87%	38.15%	64.45	51.21	42.49	—	—	—	—	—
Syntaxe	83.84%	33.33%	64.34	47.70	37.90	82.05%	38.55%	66.95	52.46	43.13
Synonymes	78.07%	35.74%	63.12	49.04	40.09	77.44%	41.37%	65.94	53.93	45.62
D	74.02%	37.75%	62.09	50.00	41.85	75.69%	43.78%	66.06	55.47	47.81
D-Syn	73.44%	37.75%	61.76	49.87	41.81	75.17%	43.78%	65.74	55.33	47.77
D-B	69.34%	38.15%	59.60	49.22	41.92	70.97%	44.18%	63.29	54.46	47.78
D-EWN-M	72.86%	40.96%	63.04	52.44	44.89	74.68%	47.39%	66.97	57.99	51.13
D-Dér	73.64%	38.15%	62.09	50.26	42.22	75.34%	44.18%	66.03	55.70	48.16
D-B-EWN-M-Dér	68.87%	41.77%	60.96	52.00	45.34	71.01%	48.19%	64.86	57.42	51.50
Tous	68.87%	41.77%	60.96	52.00	45.34	70.59%	48.19%	64.59	57.28	51.46

TAB. 7.21 – Résultats traditionnels avec et sans utilisation de la coréférence pour une réponse d'un paragraphe.

	Sans coréférence				Avec coréférence			
Enrichissement	Exact	1 exact	Faux	Sans	Exact	1 exact	Faux	Sans
Plancher	102	88	23	104	—	—	—	—
Syn. aveugles	110	94	32	97	—	—	—	—
Syntaxe	83	74	16	121	96	85	21	107
Synonymes	89	78	25	115	103	90	30	101
D	94	83	33	109	109	95	35	95
D-Syn	94	83	34	108	109	95	36	95
D-B	95	84	42	105	110	96	45	92
D-EWN-M	102	89	38	100	118	102	40	86
D-Dér	95	84	34	108	110	96	36	94
D-B-EWN-M-Dér	104	91	47	96	120	104	49	83
Tous	104	91	47	97	120	104	50	84

TAB. 7.22 – Résultats traditionnels quantitatifs comparés de l'utilisation d'un seuil de rejet des réponses sans concordance syntaxique avec la question nul ou minimal.

	Focus requis					*Focus* non requis				
Enrichissement	Précision	Rappel	F-m1	F-m2	F-m3	Précision	Rappel	F-m1	F-m2	F-m3
Plancher	83.02%	35.34%	65.38	49.58	39.93	—	—	—	—	—
Syn. aveugles	77.87%	38.15%	64.45	51.21	42.49	—	—	—	—	—
Avec coréférence	70.59%	48.19%	64.59	57.28	51.46	35.86%	57.03%	38.73	44.03	51.01
Sans coréférence	68.87%	41.77%	60.96	52.00	45.34	33.97%	50.20%	36.32	40.52	45.82

TAB. 7.23 – Résultats traditionnels de l'interrogation de la base au niveau paragraphe : variation des paramètres.

	Focus requis				*Focus* non requis			
	Exact	1 exact	Faux	Sans	Exact	1 exact	Faux	Sans
Plancher	88	77	18	118	—	—	—-	—-
Syn. aveugles	95	82	27	111	—	—	—	—
Avec coréférence	120	104	50	84	142	116	254	73
Sans coréférence	104	91	47	97	125	102	243	83

TAB. 7.24 – Résultats traditionnels quantitatifs de l'interrogation de la base au niveau paragraphe : variation des paramètres.

de réponse. Le rappel reste toutefois bien plus important lorsque ce module est exploité.

L'interrogation de la base documentaire au niveau textuel permet de confirmer nos observations sur les différentes méthodes d'enrichissement, notamment sur l'importance des synonymes contextuels, surtout s'ils proviennent d'un dictionnaire sémantique. Par ailleurs, l'importance de la présence du *focus* s'amenuise à mesure que la fenêtre de réponse est plus large car la faiblesse de la précision, qui toutefois s'est améliorée, est contrebalancée par le rappel élevé. L'importance du module de coréférence a maintenant complètement disparu, car l'antécédent anaphorique est toujours présent dans le document. Les résultats comparés de cette interrogation au niveau du document sont présentés dans les tableaux 7.25 et 7.26 page suivante. Comme les résultats de l'interrogation de la base textuelle avec utilisation de la technique de résolution de la coréférence sont identiques à ceux qui n'exploitent pas cette fonctionnalité, ces tableaux ne distinguent pas son exploitation.

	Focus requis					*Focus* non requis				
Enrichissement	Précision	Rappel	F-m1	F-m2	F-m3	Précision	Rappel	F-m1	F-m2	F-m3
Plancher	82.88%	36.95%	66.38	51.11	41.55	—	—	—	—	—
Syn. aveugles	78.86%	38.96%	65.45	52.15	43.34	—	—	—	—	—
Syntaxe	81.61%	28.51%	59.46	42.26	32.78	66.67%	32.13%	54.87	43.36	35.84
Synonymes	77.78%	30.92%	59.69	44.25	35.16	65.91%	34.94%	55.98	45.67	38.56
D	73.64%	32.53%	58.78	45.13	36.62	63.64%	36.55%	55.42	46.43	39.95
D-Syn	70.18%	32.13%	56.74	44.08	36.04	61.22%	36.14%	53.76	45.45	39.37
D-B	68.70%	31.73%	55.71	43.41	35.55	60.14%	35.74%	52.91	44.84	38.90
D-EWN-M	75.21%	36.55%	62.07	49.19	40.73	65.58%	40.56%	58.38	50.12	43.91
D-Dér	73.21%	32.93%	58.82	45.43	37.00	63.45%	36.95%	55.49	46.70	40.32
D-B-EWN-M-Dér	72.22%	36.55%	60.42	48.53	40.55	63.52%	40.56%	57.06	49.51	43.72
Tous	72.22%	36.55%	60.42	48.53	40.55	63.52%	40.56%	57.06	49.51	43.72

TAB. 7.25 – Résultats traditionnels de l'interrogation de la base au niveau texte : variation des paramètres.

Enrichissement	*Focus* requis				*Focus* non requis			
	Exact	1 exact	Faux	Sans	Exact	1 exact	Faux	Sans
Plancher	92	92	19	89	—	—	—	—
Syn. aveugles	97	97	26	77	—	—	—	—
Syntaxe	71	71	16	113	80	80	40	80
Synonymes	77	77	22	101	87	87	45	68
D	81	81	29	90	91	91	52	57
D-Syn	80	80	34	86	90	90	57	53
D-B	79	79	36	85	89	89	59	52
D-EWN-M	91	91	30	79	101	101	53	46
D-Dér	82	82	30	88	92	92	53	55
D-B-EWN-M-Dér	91	91	35	74	101	101	58	41
Tous	91	91	35	74	101	101	58	41

TAB. 7.26 – Résultats traditionnels quantitatifs de l'interrogation de la base au niveau texte : variation des paramètres.

7.4 Analyse des erreurs

À la suite de ce passage en revue des résultats de l'interrogation de la structure documentaire, nous avons essayé d'établir une typologie des difficultés récurrentes rencontrées par notre système pour mettre en correspondance les questions et leurs réponses. En effet , il paraît intéressant d'identifier les causes de dysfonctionnement du système afin de pouvoir les pallier à l'avenir.

7.4.1 Erreurs liées aux ressources lexicales

Pour la création de la structure informationnelle, nous exploitons une information essentiellement extraite de dictionnaires. Cependant, les imperfections peuvent se glisser dans la structure informationnelle soit du fait de l'information elle-même, qui n'est pas irréprochable, soit à cause d'une exploitation incomplète de l'information présente. Nous avons décelé trois type d'imperfections récurrentes dans l'utilisation des ressources lexicales. Elles sont liées à la synonymie, à la dérivation morphologique ou aux verbes auxiliaires[18].

[18]Nous utilisons ici le terme *auxiliaire* au sens qu'il prend en grammaire générative, c'est-à-dire « une catégorie grammaticale constituant obligatoire du syntagme verbal et comprenant un constituant de temps, d'aspect et de modalité » [Dubois et al., 1999]. [Grevisse et Goosse, 1991] désignent ces verbes sous le terme « **semi-auxiliaires** ».

Synonymie

Dans de nombreux cas, des paires de synonymes qui paraissent évidentes ne sont pas connues par les dictionnaires. Il est vrai que nous avons souligné la médiocrité du dictionnaire de synonymes *Bailly* dont nous disposons. Par ailleurs, l'information extraite des dictionnaires sémantiques est restreinte et la synonymie ne constitue dans *EuroWordNet* qu'une relation sémantique étudiée parmi d'autres. Il serait donc intéressant de pouvoir disposer d'une information riche, fiable et spécifiquement synonymique pour améliorer cette phase d'enrichissement de la structure informationnelle. Nous avons relevé plus de quinze cas où une réponse ne pouvait être mise en rapport avec la question du fait d'une synonymie inconnue. Par exemple *gendre* n'est pas mis en relation avec *beau-fils* dans les dictionnaires dont nous disposons, alors qu'il est parfaitement synonyme d'au moins une de ses acceptions.

Dans la perspective d'un approfondissement de l'utilisation de la synonymie dans une structure où la syntaxe joue un rôle important, il est également important d'identifier les variations syntaxiques et syntaxico-sémantiques du contexte dans le cas où l'on considère, comme nous l'avons fait, qu'un synonyme remplace le mot de départ dans la phrase pour y former un nouvel énoncé de même sens.

$$X \textit{ est fatal à } Y \approx Y \textit{ meurt de } X$$

Certains cas, proches de la synonymie, concernent plutôt l'instauration d'une hiérarchie hypéronymique ou holonymique qui permettrait de mettre en rapport deux termes contenus l'un dans une question et l'autre dans sa réponse à travers la généralisation de l'un ou la spécialisation de l'autre. En effet, dans plusieurs cas, le texte est plus précis dans l'emploi des termes que la question posée à son propos. Par exemple, une des questions posées demandait *Quelle était la fonction de Marc Antoine en 44 ?* tandis que le texte dit que *Marc Antoine devint consul en 44*. Le fait de savoir que la charge de *consul* est une *fonction* permettrait de mettre en rapport la question et sa réponse.

Dès la description des ressources lexicales, nous avons mis en évidence le caractère sémantique hiérarchique de *EuroWordNet* et le choix d'utiliser ce dictionnaire sémantique dans notre démarche a été déterminé autant pour ses possibilités généralisatrices que pour la richesse synonymique qu'elle a été capable de nous apporter.

Dérivation morphologique

Les remarques que nous avons faites sur les résultats décevants de l'enrichissement au travers de dérivés morphologiques sont justifiées. En effet,

nous avons rencontré onze cas où l'utilisation de verbes dans la question ne pouvait amener à une mise en correspondance de l'information contenue dans cette question et de celle de la réponse correcte où l'entité de même sens appartenait à une autre catégorie grammaticale.

> *Qui* ***protégeait*** *Suétone ?*
> *Lorsque meurt Pline le Jeune,* ***protecteur*** *de Suétone, Septicius, préfet du prétoire, introduit Suétone à la cour, lui permettant d'aborder une grande carrière publique sous le règne d'Hadrien.*

Dans l'exemple, *protecteur* présent dans le texte ne trouve pas dans le *Dubois* l'information de dérivation qui permettrait d'établir son rapport avec *protéger* présent dans la question. Ce verbe possède pourtant l'information de dérivation qui permettrait de retrouver *protecteur* au départ de *protéger*. Or cette information n'est pas utilisée au départ de la question, dont le contexte n'est généralement pas suffisant pour assurer une désambiguïsation sémantique correcte et, de ce fait, une sélection des autres informations lexicales. Par ailleurs, certaines parentés entre noms sont signalées dans un sens et pas dans l'autre. Ainsi, *biographe* est relié à *biographie*, mais l'inverse n'est pas vrai.

Nous pensons donc qu'il serait judicieux d'approfondir ce type d'enrichissement. Ce peut être réalisé en utilisant une ressource qui permettrait d'établir une connexion morphologique et sémantique entre chaque sens de chaque unité lexicale et l'ensemble des mots qui en dérivent ou dont cette unité lexicale est elle-même le dérivé dans le sens étudié. Il est aussi possible de recourir à un enrichissement de l'information de la partie générale du *Dubois* à l'aide de sa partie verbale d'une manière semblable à celle que nous avons pratiquée pour la contextualisation des synonymes, en exploitant les domaines d'application.

(Semi-)auxiliaires

Les seuls auxiliaires identifiés par l'analyse morpho-syntaxique sont les unités verbales fonctionnelles qui servent à façonner les formes verbales composées (*avoir* et *être*). À ce titre, ils sont automatiquement exclus des dépendances significatives construites par l'analyseur et laissent la place au lemme du verbe à ses formes simples. Il y a pourtant une catégorie de « verbes qui, construits avec un infinitif, parfois avec un participe ou un gérondif, perdent plus ou moins leur signification propre et servent à exprimer diverses nuances de temps, d'aspect ou d'autres modalités de l'action » [Grevisse et Goosse, 1991], § 789. Or la présence dans l'énoncé de requêtes de ces auxiliaires, qui dans d'autres schémas syntaxiques possèdent généralement un tout autre sens, contrarie ordinairement l'identification de la

réponse dans laquelle cet auxiliaire n'est pas présent.

Où ***se trouvent*** *les champs Décumates ?*
Qu'est-il ***arrivé*** *à la famille de Julien lorsqu'il avait six ans ?*

La présence des semi-auxiliaires *se trouver* et *arriver* dans ces exemples de questions perturbe le bon fonctionnement de l'appariement question-réponse, car le système réclame dans les propositions de réponse respectivement *trouver* et *arriver*, qui ont bien peu de chances de s'y trouver (*[...] l'annexion des champs Décumates, territoires compris entre les cours supérieurs du Rhin et du Danube, a pour but [...]* et *[...] dès l'âge de six ans, après le massacre de sa famille, ordonné par les successeurs de Constantin (337), Julien [...]*). Les deux cas présentés ne trouvent pas de semi-auxiliaire semblable dans la réponse et dans la question.

Ici encore, l'information présente dans le dictionnaire *Dubois* permet d'identifier ces lexèmes comme des auxiliaires sous certaines de leurs acceptions. Notre décision de ne pas effectuer de désambiguïsation sémantique sur les énoncés de questions ne permet pas toutefois de décider si le verbe est utilisé dans une de ces acceptions ou non. Il est toutefois possible de créer un nouveau cas de relâchement de contrainte comme celui que nous avons imaginé pour l'unité lexicale désignée par le *focus*, mais qui porterait cette fois sur les lexèmes verbaux dont une des acceptions le décrit comme un auxiliaire. Ce type de dégradation informationnelle de la requête permettrait d'éliminer le lexème des informations requises dans les réponses dans les cas où aucune réponse n'a été trouvée pour la question.

7.4.2 Erreurs liées à l'analyse du texte ou de la question

Nous avons noté au cours de l'évaluation du système que certains défauts récurrents provenaient de défauts de l'analyse appliquée aux documents ou à la question. Ils peuvent dépendre de trois niveaux d'analyse : morphologique, syntaxique ou résolution des coréférences.

Analyse morphologique

Dans un corpus comme celui que nous avons utilisé pour évaluer la pertinence de notre démarche, les noms propres qui sont entrés dans la langue comme noms communs sont très nombreux (*césar*, *auguste*, *commode*, *galère* etc.). Par ailleurs, certains autres noms propres ne sont simplement pas recensés dans les lexiques, et leur identification est problématique. *Pompée*, par exemple, sera plus souvent identifié comme une forme fléchie de *pomper* que comme le nom d'un général romain (cinq fois sur sept apparitions dans

le corpus de questions). Ces confusions suscitent des erreurs dans l'extraction des dépendances sur lesquelles est fondée la structure informationnelle. Dès lors, certaines mises en correspondance de réponses avec la question ne peuvent s'effectuer.

Outre les noms propres, certains autres lexèmes ne sont pas identifiés correctement par l'analyse morphologique et la désambiguïsation catégorielle. Nous avons en effet décelé plusieurs cas de distinction douteuse entre nom et adjectif. Le mot *partisan* par exemple peut être nom ou adjectif. Cependant, la seule acception adjectivale de *partisan* dans le *Dubois* correspond à *partial*, ce qui provoque des erreurs d'enrichissement dans les cas où la catégorie adjectivale est sélectionnée par la désambiguïsation catégorielle et que cette acception n'est pas correcte. D'autre part, les lexèmes *partisan* nominal et adjectival ne coïncident pas dans notre méthode, et ne peuvent donc permettre la mise en correspondance de réponses avec la question.

Une erreur de désambiguïsation catégorielle est donc virtuellement à l'origine de nombreuses sources d'erreur : mauvaise correspondance des lexèmes présents dans la structure informationnelle, analyse syntaxique incorrecte, désambiguïsation sémantique fautive et donc enrichissements tout aussi fautifs. Une semblable erreur dans l'analyse d'une question ne résout pas le problème car nous ne pouvons prévoir que son analyse syntaxique erronée correspondra à celle qui a été effectuée dans le document. Par exemple, la question *Qui l'usurpateur Magnence a-t-il assassiné ?* ne peut être appariée à sa réponse *Il périt assassiné par l'usurpateur Magnence.* En effet, l'étiquetage du mot *assassiné* comme adjectif est fautif et il ne permet pas d'identifier *usurpateur Magnence* comme un complément d'agent. Dès lors, la mise en rapport du sujet d'un verbe actif ne peut être effectuée avec l'agent du même verbe passif.

Dépendances syntaxiques

Les erreurs d'analyse syntaxique qui ne sont pas issues d'un étiquetage erroné des unités lexicales sont relativement nombreuses, surtout en ce qui concerne la partie du traitement consacrée aux questions. En effet, la grammaire française que nous utilisons est expérimentale. De plus, elle a été écrite pour gérer du texte tout venant. Les questions sont relativement rares dans les textes utilisés pour construire les grammaires, qui sont généralement extraits de divers types d'articles de la presse. De plus, les questions présentes dans ces textes sont souvent oratoires et ne présentent donc pas forcément les mêmes caractéristiques que des questions réelles comme celles que nous avons à traiter dans le cadre de cette évaluation. Quoi qu'il en soit, l'analyseur rencontre des difficultés pour gérer les dépendances de base comme le sujet (`SUBJ`) ou l'objet (`VARG[DIR]` et `VARG[INDIR]`) à cause de la structure

de la phrase, inversée fréquemment ou très particulière à cause de l'interrogation : *-t-* épenthétique (*Où Marius a-**t**-il été fait prisonnier ?*), répétition du sujet (*Qui **Nerva** eut-**il** pour consul ?*), particularités interrogatives (*Quand **est-ce que** Théodose est mort ?*), etc.

D'autres erreurs apparaissent dans l'analyse des documents eux-mêmes : nous avons par exemple noté des erreurs lorsque un nom est composé (*Il fait édifier une muraille continue d'une mer à l'autre, le **vallum Hadriani***) ou lorsque les composantes d'une expression verbale sont très éloigné l'une de l'autre (*Septime Sévère **fut**, à la mort de Pertinax (193), **proclamé** empereur par les légions d'Illyrie*). Nous n'insistons pas sur les erreurs de rattachement prépositionnel ou de coordination, bien connues en analyse syntaxique automatique et loin d'être triviales à résoudre. D'autres dépendances sont simplement absentes.

Résolution d'anaphore

Nous avons suffisamment insisté sur l'importance de la résolution de la coréférence au cours de cette évaluation. La technique que nous avons adoptée, pour grossière qu'elle est, ne suffit pas moins à montrer la prépondérance de ce type de lien dans les applications de gestion de l'information à un niveau inférieur au texte. Nous n'avons pas manqué toutefois de signaler les limites de la méthode de résolution de coréférence utilisée : non seulement elle n'est pas extensible à d'autres types de corpus – et même dans le cas présent elle commet un nombre important d'erreurs qui ne sont pas directement sensibles dans les résultats de l'interrogation étant donné que l'apport de réponses exactes surclasse de beaucoup le nombre de réponses inexactes fournies – mais encore elle ne peut s'appliquer qu'aux pronoms personnels sujets.

Nous avons rencontré de nombreux cas où la résolution de coréférence de pronoms autres que les personnels sujets ou d'adjectifs possessifs permettrait la mise en correspondance de réponses avec la question. En effet, dans la phrase *Octavien exploita l'indignation (...) pour abattre **son** rival*, l'identification de l'entité déterminée par *son* à *Octavien* permettrait une correspondance de *rival d'Octavien* avec *son rival*. La grammaire de [Trouilleux, 2001] permet de relier adjectifs possessifs et pronoms avec leur coréférent. Son adaptation à la version de l'analyseur syntaxique que nous utilisons n'a été effectuée qu'après que nous avons mené cette évaluation et nous n'avons donc pas pu en exploiter les capacités.

7.4.3 Erreurs liées à un besoin de logique ou de connaissances du monde

Viennent ensuite diverses constatations d'erreurs ou de silences dus à des phénomènes sans rapport direct avec la linguistique, mais qui font appel au bon sens, au jugement ou à la pragmatique. Dans les spécifications que nous avons édictées sur le protocole d'évaluation établi et suivi, nous avons demandé aux utilisateurs de ne pas faire appel à un jugement ou à une déduction de la part du système. Certaines questions ont toutefois transgressé ces spécifications, alors que les utilisateurs connaissaient les textes. Il nous paraît donc évident que limiter un système de gestion de l'information selon ce type de critère est abusif. Nous avons dès lors tenté d'identifier certaines carences actuelles de notre système au niveau de la logique et de la pragmatique afin de déterminer d'éventuelles solutions.

Dans de nombreux cas, nous avons trouvé des questions pour lesquelles seule une déduction logique permettrait leur mise en correspondance avec le fragment de texte qui en constitue la réponse. Nous avons ainsi décelé des inversions de liens sémantique (*Qui est le* ***père de*** *Caracalla ? – Caracalla est le premier* ***fils de*** *Septime Sévère*), des implications (*Quels étaient les adversaires de Julien lors de la* ***bataille*** *de Strasbourg ? – Il remporte sur les Alamans l'éclatante* ***victoire*** *de Strasbourg*) ou des déductions logiques (*Quel mois de l'année a été nommé* ***en hommage à*** *César ? – Le mois de sa naissance est nommé « juillet »* [19]).

D'autres questions font appel au bon sens ou à la connaissance du monde. Ainsi, le véritable nom de Germanicus est Julius Caesar [20]. Aussi, lorsque la question *Quel titre Julius Caesar se donne-t-il ?* est posée, le système éprouve-t-il des difficultés à faire un choix entre les différents *Caesar* qui lui sont proposés. De même une connaissance du monde approfondie est-elle nécessaire pour apporter à la question *Quel est le surnom de Metellus* l'extrait de texte *Caecilius Metellus, dit le Macédonique*, ou à la question *En quelle année Julien est-il proclamé empereur ?* le fragment *En 360, ses soldats se mutinent et le proclament auguste* [21]. La pragmatique et des connaissances générales plus ou moins approfondies entrent ici en jeu.

Cependant, un certain nombre de liens logiques peuvent être établis entre des unités lexicales de même catégorie grammaticale grâce à des ressources de type *EuroWordNet* dont les relations sémantiques permettent certains types d'inférence. Il en est ainsi du lien de filiation qui s'inverse selon qu'il

[19] Cet exemple illustre également l'importance de la résolution de la coréférence portant sur l'adjectif possessif *sa* dont il faut savoir qu'il renvoie à *Jules César*.

[20] C'est-à-dire le même nom que celui du Jules César que nous connaissons.

[21] C'est Dioclétien qui en 286 instaura la tétrarchie, système de pouvoir où deux augustes se partagent le pouvoir (Occident et Orient) en tant qu'empereurs, assistés par deux césars qui doivent en principe leur succéder.

est actualisé par *fils de* ou par *père de*. Une autre inférence, l'implication, relie le fait qu'il y ait *victoire* à Strasbourg et le fait qu'une *bataille* s'y est déroulée. Ces liens logiques peuvent dans certains cas pallier les manques logiques d'une approche purement linguistique.

Conclusion

Dans une société où l'information a pris une importance vitale, la maîtrise des données contenues dans les documents électroniques est devenue un enjeu capital. Cependant, la gestion de l'information électronique, chaque jour plus volumineuse, n'est envisageable qu'avec l'aide de techniques automatiques, auxquelles on demande de trier, de classer, de filtrer, d'extraire ou d'interroger l'information en fonction des besoins propres à chaque utilisateur.

Les méthodes de gestion automatique de l'information se heurtent généralement à deux obstacles importants. Le premier de ces obstacles découle du grand nombre de possibilités d'actualisations d'une même information en langage naturel. Ce problème est traité de diverses manières par les méthodologies du domaine : synonymie, racinisation, constitution de matrices et de lexiques à partir de corpus étiquetés, etc. La plupart du temps, ces approches recourent à une spécialisation du domaine pour restreindre les possibilités de diversification de la forme ainsi que du type d'information à identifier, ou à des mesures statistiques qui tiennent peu compte du contenu réel des textes.

La seconde difficulté concerne la compréhension des énoncés afin d'identifier la pertinence d'une information dans le cadre du traitement considéré. L'identification de cette pertinence est habituellement effectuée par un calcul de similarité de vecteurs ou de matrices censés représenter l'information, ou par l'évaluation de la capacité d'une information à rentrer dans un tableau informatif précis. Ici encore, les solutions proposées sont statistiques et donc se penchent peu sur le contenu des textes, ou elles restreignent le domaine d'application de la méthodologie.

Nous constatons toutefois qu'au cours de l'histoire de l'élaboration des méthodes proposées pour la gestion de l'information, elles intègrent de plus en plus d'éléments linguistiques, d'abord des ressources lexicales, ensuite des outils de traitement, pour identifier et mettre en relation les différentes unités qui composent les textes. Ainsi, les systèmes actuels utilisent souvent une ou plusieurs de ces techniques : découpage en mots et normalisation des unités lexicales, analyse morphologique, constitution de syntagmes, établissement de relations syntaxiques, ébauches de traitement sémantique. L'apport de ces

outils linguistiques est souvent constaté, mais reste généralement générique dans les méthodes existantes.

L'amélioration des performances et de la robustesse des analyseurs linguistiques permet actuellement d'envisager leur utilisation dans des applications réelles. Nous proposons donc d'utiliser exclusivement des approches de type linguistique pour construire un système de structuration de l'information générique qui permettra de manipuler une information dont le sens a été identifié. Cette méthodologie a pour objet d'étudier et d'identifier les contenus de documents sans restriction de domaine pour permettre leur maniement dans les différentes perspectives de la gestion de l'information.

Pour réaliser ce système, nous nous sommes appuyé sur une réflexion existante qui portait sur deux points : d'abord, l'enrichissement (aussi appelé expansion) d'un énoncé permet de donner un grand nombre d'actualisations différentes à un même énoncé ; ensuite, plus le contexte d'un mot est riche et précis, plus l'identification du sens de ce mot en est facilitée. Or les différents secteurs de la gestion de l'information s'appuient généralement sur une constatation de similitude entre deux informations. En conséquence, il nous a semblé judicieux d'effectuer l'identification du sens des unités lexicales à l'intérieur des documents dont le contexte est généralement plus riche que celui d'une requête, et d'enrichir ensuite chaque énoncé des documents sur la base des sens identifiés.

La construction d'une structure informationnelle s'appuie dès lors sur une analyse linguistique aussi complète que possible, c'est-à-dire l'identification des mots, l'analyse morphologique et l'établissement des relations syntaxiques. Cette analyse est nécessaire pour effectuer la désambiguïsation sémantique qui identifie le sens des unités lexicales en contexte. Par ailleurs, les liens syntaxiques sont susceptibles de permettre l'établissement de relations syntaxico-sémantiques (par exemple actant-action, action-patient, etc.). L'identification des schémas syntaxiques et du sens des mots des énoncés permet ensuite d'effectuer des enrichissements dont l'originalité est d'être soumis à l'identification des sens des mots et des schémas syntaxiques des phrases.

Les différents types d'enrichissements sont issus de techniques déjà testées dans le domaine : synonymie simple et expressions synonymiques, classes et catégories sémantiques, domaines d'application, dérivation morphologique. Toutes les données permettant d'enrichir les textes proviennent de ressources lexicales et lexico-sémantiques, et sont identifiées non au départ du lexème, mais à partir de son sens. De plus, l'utilisation d'unités lexicales dans l'enrichissement est basée sur le principe de l'interchangeabilité. En effet, ces enrichissements sont apportés de manière à ce que chaque lexème issu de l'enrichissement remplace dans l'énoncé le lexème qu'il enrichit et modifie éventuellement la structure syntaxique de manière à conserver un énoncé

correct et sémantiquement équivalent à l'énoncé de départ.

En plus de ces enrichissements, une méthode simple permet d'identifier le coréférent des pronoms personnels sujets présents dans les textes de la base documentaire.

La structure informationnelle est un index des unités lexicales présentes dans le texte ou des lexèmes correspondant à leurs enrichissements reliées entre eux par des dépendances syntaxico-sémantiques. Cet index permet de retrouver dans la base documentaire les textes ou fragments de textes correspondant à une information déterminée.

Pour tester la qualité de la structure informationnelle, nous avons élaboré un module d'interrogation qui permet de l'interroger à trois niveaux : texte, paragraphe et phrase. L'interrogation s'effectue à partir de requêtes en langage naturel, qui sont analysées semblablement aux documents, mais ne sont ni désambiguïsées, ni enrichies car le contexte de ces requêtes est généralement trop pauvre pour permettre ce type de traitement. L'information obtenue à partir des résultats de l'analyse est comparée à l'information de la structure informationnelle pour obtenir les fragments de texte correspondants.

Divers traitements peuvent être appliqués à la requête pour assouplir les contraintes de correspondance, très élevées au départ, des réponses candidates avec la question. Ces contraintes portent essentiellement sur l'objet de la question et sur le taux de correspondance des dépendances syntaxiques entre la question et la réponse candidate.

L'évaluation que nous avons effectuée a porté sur deux types de tâches très exigeantes de la gestion de l'information. Il s'agit de la tâche de question-réponse, et d'un calcul de résultats se rapprochant de l'extraction d'information. Les résultats que nous avons obtenus dans les deux cas sont très honorables, et démontrent l'intérêt de l'identification du sens de unités lexicales qui constituent les documents pour effectuer un enrichissement contextuel. Dans la tâche de question-réponse, le fonctionnement du système avec un minimum de contraintes obtient le meilleur résultat car la classification des réponses permet d'éliminer les réponses candidates trop éloignées de la question. Dans le calcul des résultats en terme de précision et de rappel, le maintien de certaines contraintes (absence du lexème désigné par le *focus* et identification sémantique de l'objet de la requête, élimination des réponses dont les dépendances syntaxiques ne concordent qu'à moins de 20 % avec celles de la requête) permet d'obtenir le meilleur compromis entre la précision (78,29 %) et le rappel (44,89 %).

La qualité de l'information fournie par la dérivation morphologique se révèle excellente, mais elle est peu utilisée. L'importance de la résolution d'anaphores est capitale dans tous les cas. Par ailleurs, les traitements aveugles,

comme l'enrichissement brut des unités lexicales dont le sens n'a pas été identifié, sont inutiles. De plus, les traitements relativement fins que nous appliquons sont moins déterminants à mesure que la fenêtre de réponse s'élargit. Par exemple, la résolution de coréférence n'améliore pas les résultats d'une interrogation au niveau du texte.

L'étude des résultats et des silences de la méthode permet d'identifier certaines de ses faiblesses et de proposer des améliorations au système. Ces amélioration sont de deux ordres :

Au niveau de l'enrichissement :

- il est d'abord important de trouver d'autres ressources synonymiques qui permettraient d'augmenter le volume de l'enrichissement ;
- ensuite, il s'agit d'accroître les possibilités de dérivation morphologique à partir des noms et des adjectifs pour accorder à ces catégorie grammaticale une information semblable à celle des verbes ;
- de plus, l'identification des entités (humain, date, lieu, etc.) et des noms propres (Pompée, Commode) est de nature à éliminer de nombreuses erreurs d'analyse et d'enrichissement ;
- enfin, l'intégration d'une grammaire de résolution des coréférences est déterminante pour le bon fonctionnement du système. L'adaptation de la méthode de [Trouilleux, 2001] est en cours.

Au niveau de l'interrogation :

- l'identification des semi-auxiliaires est un point important qui permettrait d'éliminer une grande partie des silences auxquels la méthode est actuellement confrontée ;
- l'utilisation de relations sémantiques jusque maintenant inexploitées devrait permettre la mise en correspondances d'unités lexicales différentes mais décrivant des réalités liées, soit par des liens de type hyponymiques ou holonymiques, soit par des liens logiques (inférence, implication, cause) que la ressource *EuroWordNet* est capable de fournir dans une certaine mesure ;
- enfin, le développement de règles syntaxiques adaptées aux questions devrait également éliminer bien des erreurs d'analyse, et donc supprimer des réponses inexactes et amener des réponses correctes.

Par ailleurs, la disponibilité de ressources plus nombreuses et plus complètes pour l'anglais, ainsi que d'outils souvent plus aboutis rendrait intéressante l'adaptation de notre méthodologie à cette langue. En particulier, la richesse de la ressource *WordNet* et la mise à disposition de corpus étiquetés selon les sens qu'il décrit permet d'envisager un processus d'analyse et d'enrichissement fondé entièrement sur *WordNet* et qui exploiterait au maximum les différentes relations sémantiques décrites.

Bibliographie

[Aho et Ullman, 1973] Aho, Alfred V. et Ullman, Jeffrey D., *The Theory of Parsing, Translation and Compiling*, Prentice-Hall, 1973.

[Aho et al., 1988] Aho, Alfred V., Sethi, Ravi et Ullman, Jeffrey D., *Compilers. Principles, Techniques and Tools*, Addison-Wesley, 1988.

[Appelt, 1999] Appelt, Douglas E., *Introduction to information extraction*, dans *Artficial Intelligence Communications*, tm. 12(3), 1999, pp. 161–172.

[Atkins, 1993] Atkins, Sue, *Tools for Computer-aided Corpus Lexicography : the Hector Project*, dans *Acta Linguistica Hungarica*, tm. 41, 1993, pp. 5–72.

[Aït-Mokhtar, 1998] Aït-Mokhtar, Salah, *L'analyse présyntaxique en une seule étape*, Thèse de doctorat, Université Clermont 2 Blaise Pascal, Clermont-Ferrand, 1998.

[Aït-Mokhtar et Chanod, 1997a] Aït-Mokhtar, Salah et Chanod, Jean-Pierre, *Incremental Finite-State Parsing*, dans *Proceedings of ANLP'97*, 1997a, pp. 72–79.

[Aït-Mokhtar et Chanod, 1997b] Aït-Mokhtar, Salah et Chanod, Jean-Pierre, *Subject and Object Dependecy Extraction Using Finite-State Transducers*, dans *ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications*, 1997b, pp. 71–77.

[Aït-Mokhtar et al., 2002] Aït-Mokhtar, Salah, Chanod, Jean-Pierre et Roux, Claude, *Robustness beyond shallowness : incremental deep parsing*, dans *Natural Language Engineering*, tm. 8(2/3), 2002, pp. 121–144.

[Bailly et Toro, 1947] Bailly, René et Toro, André, *Dictionnaire des synonymes de la langue française*, Larousse, Paris, 1947.

[Black, 1988] Black, Ezra, *An experiment in computational discrimination of english word senses*, dans *IBM Journal of Research and Development*, tm. 32(2), 1988, pp. 185–194.

[Bookman, 1987] Bookman, Lawrence A., *A Microfeature Based Scheme for Modelling Semantics*, dans *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 1987, pp. 611–614.

[Briscoe, 1991] Briscoe, Ted, *Lexical Issues in Natural Language Processing*, dans Klein, Ewan H. et Veltman, Frank (réds.), *Natural Language and Speech*, Springer-Verlag, 1991, pp. 39–68.

[Brun, 2000] Brun, Caroline, *A Client/Server Architecture for Word Sense disambiguation*, dans *Proceedings of Coling'2000*, Saarbrücken, Deutschland, 2000, pp. 132–138.

[Brun et al., 2001] Brun, Caroline, Jacquemin, Bernard et Segond, Frédérique, *Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale*, dans *Traitement Automatique des Langues*, tm. 42(3), 2001, pp. 667–690.

[Califf et Mooney, 1997] Califf, Mary Elaine et Mooney, Raymond J., *Relational Learning of Pattern-Match Rules for Information Extraction*, dans *Working Papers of ACL-97 Workshop on Natural Language Learning*, ACL'97, 1997, pp. 9–15.

[Califf et Mooney, 1999] Califf, Mary Elaine et Mooney, Raymond J., *Relational Learning of Pattern-Match Rules for Information Extraction*, dans *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, AAAI-99, 1999, pp. 328–334.

[Calzolari et Corazzari, 2000] Calzolari, Nicoletta et Corazzari, Ornella, *Senseval/Romanseval : The Framework for Italian*, dans *Computer and the Humanities. Special Issue on SENSEVAL*, tm. 34(1-2), 2000, pp. 61–78.

[Catherin, 1999] Catherin, Laurent, *The French WordNet*, Rap. tech. Deliverable 2D014, EuroWordNet, 1999.

[Chinchor, 1992] Chinchor, Nancy, *MUC-4 Evaluation Metrics*, dans *Proceedings of the Fourth Message Understanding Conference*, MUC-4, Morgan Kaufmann, San Mateo, 1992, pp. 22–29.

[Chinchor et al., 1994] Chinchor, Nancy, Hirschman, Linette et Lewis, David D., *Evaluating Message Understanding Systems : An Anlysis of the Third Message Understanding Conference (MUC-3)*, dans *Computational Linguistics*, tm. 19(3), 1994, pp. 409–449.

[Church, 1995] Church, Kenneth Ward, *One term or two ?*, dans *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM SIGIR, Seattle, Washington, United States, 1995, pp. 310–318.

[Corréard et Grundy, 1994] Corréard, Marie-Hélène et Grundy, Valérie (réds.), *The Oxford-Hachette French Dictionary*, Oxford University Press, Hachette, Oxford, Paris, 1994.

[Cowie et al., 1992] Cowie, Jim, Guthrie, Joe A. et Guthrie, Louise, *Lexical disambiguation using simulated annealing*, dans *Proceedings of the 14th International Conference on Computational Linguistics*, 1992, pp. 23–28.

[Dagan et al., 1991] Dagan, Ido, Itai, Alon et Schwall, Ulrike, *Two lnguages are more informative than one*, dans *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, 1991, pp. 130–137.

[Dagan et al., 1993] Dagan, Ido, Marcus, Shaul et Markovitch, Shaul, *Contextual word similarity and estimation from sparse data*, dans *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, 1993, pp. 164–171.

[Dahlgren, 1988] Dahlgren, Kathleen, *Naive Semantics for Natural Language Understanding*, Kluwer Academic Publishers, New York, 1988, 258 pp.

[Dini et al., 1998] Dini, Luca, Di Tomaso, Vittorio et Segond, Frédérique, *Error Driven Word Sense Disambiguation*, dans *Proceedings of the Conference COLING-ACL'98*, COLING-ACL, Montréal, 1998, pp. 320–324.

[Dini et al., 2000] Dini, Luca, Di Tomaso, Vittorio et Segond, Frédérique, *GINGER II : an example-driven word sense disambiguator*, dans *Computer and the Humanities. Special Issue on SENSEVAL*, tm. 34(1-2), 2000, pp. 121–126.

[Dubois, 1967] Dubois, Jean, *Grammaire structurale du français. T.2 Le verbe*, Larousse, Paris, 1967.

[Dubois et Dubois-Charlier, 1997] Dubois, Jean et Dubois-Charlier, Françoise, *Dictionnaire des verbes français*, Larousse, Paris, 1997. La première version de ce dictionnaire est électronique. Elle est accompagnée de son complément *Dictionnaire des mots*.

[Dubois et al., 1999] Dubois, Jean, Giacomo, Mathée, Guespin, Louis, Marcellesi, Christiane, Jean-Baptiste, Marcellesi et Mével, Jean-Pierre, *Dictinnaire de linguistique et des sciences du langage*, Larousse-Bordas, Paris, 1999.

[Dyer, 1983] Dyer, M. (réd.), *In-depth Understanding*, MIT Press, Cambridge, Massachusets, 1983.

[Fellbaum, 1990] Fellbaum, Christiane, *English verbs as a semantic net*, dans *International Journal of Lexicography*, tm. 3(4), 1990, pp. 278–301.

[Fellbaum, 1998a] Fellbaum, Christiane, *Semantic Network of English Verbs*, dans Fellbaum, Christiane (réd.), *WordNet : an electronic lexical database*, The MIT Press, Cambridge, Massachusetts, 1998a, pp. 69–104.

[Fellbaum, 1998b] FELLBAUM, Christiane (réd.), *WordNet : an electronic lexical database*, Language, Speech and Communication, The MIT Press, Cambridge, Massachusetts, 1998b.

[Ferret et al., 1999] FERRET, Olivier, GRAU, Brigitte, ILLOUZ, Gabriel, JACQUEMIN, Christian et MASSON, N., *QUALC. The question-answering program of the Langage et Cognition group at LIMSI-CNRS*, dans *Proceedings of The Eighth Text Retrieval Conference*, 1999, pp. 455–464.

[Ferret et al., 2002a] FERRET, Olivier, GRAU, Brigitte, HURAULT-PLANTET, Martine, ILLOUZ, Gabriel et JACQUEMIN, Christian, *Quand la réponse se trouve dans un grand corpus*, dans *Revue d'Ingénierie des Systèmes d'Information*, tm. 7(1-2), 2002a, pp. 95–123.

[Ferret et al., 2002b] FERRET, Olivier, GRAU, Brigitte, HURAULT-PLANTET, Martine, ILLOUZ, Gabriel, MONCEAUX, Laura, ROBBA, Isabelle et VILNAT, Anne, *Recherche de la réponse fondée sur la reconnaissance du focus de la question*, dans *Actes de TALN 2002*, 2002b, pp. 98–107.

[Freitag, 1998] FREITAG, Dayne, *Multistrategy Learning for Information Extraction*, dans *Proceedings of the Fifteenth International Machine Learning Conference*, 1998, pp. 161–169.

[Fukumoto et Kato, 2001] FUKUMOTO, Jun'ichi et KATO, Tsuneaki, *An overview of Question and Answering Challenge (QAC) of the Next NTCIR Workshop*, dans *Proceedings of the 2nd NTCIR Workshop on Research on Chinese and Japanese Text Retrieval and Text Summarisation*, NII, 2001.

[Fukumoto et al., 2003] FUKUMOTO, Jun'ichi, KATO, Tsuneaki et MASUI, Fumiko, *Question Answering Challenge (QAC-1) : An Evaluation of Question Answering Task at NTCIR Workshop 3*, dans *Proceedings of the 3rd NTCIR Workshop on Research on Chinese and Japanese Text Retrieval and Text Summarisation*, NII, 2003.

[Gale et al., 1992] GALE, William A., CHURCH, Kenneth W. et YAROWSKY, David, *One sense per discourse*, dans *Proceedings of the Speech and Natural Language Workshop*, Morgan Kaufmann, 1992, pp. 233–237.

[Gale et al., 1993] GALE, William A., CHURCH, Kenneth W. et YAROWSKY, David, *A method for disambiguating word senses in large corpus*, dans *Computer and the Humanities*, tm. 26, 1993, pp. 415–439.

[Gaussier, 1999] GAUSSIER, Éric, *Unsupervised learning of derivational morphology from inflectional lexicons*, dans *ACL'99 Workshop Proceedings on Unsupervised Learning in Natural Language Processing*, ACL'99, College Park, Maryland, USA, 1999, pp. 24–30.

[Gaussier et al., 1997] GAUSSIER, Éric, GREFENSTETTE, Gregory, HULL, David et SCHULZE, B. Maximilian, *Xerox TREC-6 Site Report : Cross*

Language Text Retrieval, dans *Proceedings of The Sixth Text Retrieval Conference*, TREC-6, 1997, pp. 775–788.

[Gaussier et al., 2000] Gaussier, Éric, Grefenstette, Gregory, Hull, David et Roux, Claude, *Recherche d'information en français et traitement automatique des langues*, dans *Traitement Automatique des Langues*, tm. 41(2), 2000, pp. 473–493.

[GENELEX, 1994] GENELEX, Consortium, *Projet EUREKA GENELEX. Rapport sur la couche sémantique*, ASSTRIL, Gsi-Erli, IBM France, SEMA GROUP, 2e édn., 1994.

[Graesser et al., 1994] Graesser, A.C., McMahen, C.L. et Johnson, B.K., *Question Asking and Answering*, dans Gernsbacher, Morton Ann (réd.), *Handbook of Psycholinguistics*, Academic Press, 1994, pp. 517–538.

[Grevisse et Goosse, 1991] Grevisse, Maurice et Goosse, André, *Le bon usage*, Duculot, Paris, Louvain-la-Neuve, 1991, 12e édn.

[Habert, 2000] Habert, Benoît, *Création de dictionnaires sémantiques et typologie des textes*, dans Tyvaert, Jean-Emmanuel (réd.), *Actes des journées scientifiques 1999. L'Imparfait – Philologie électronique et assistance à l'interprétation des textes*, 15, CIRLEP (Centre Interdisciplinaire de Recherches en Linguistique et Psychologie cognitive, Presses Universitaires de Reims, Reims, 2000, pp. 171–188.

[Hagège et Roux, 2002] Hagège, Caroline et Roux, Claude, *A Robust and Flexible Platform for Dependency Extraction*, dans *Proceedings of LREC 2002*, LREC 2002, Las Palmas, Canaria, España, 2002, pp. 520–523.

[Hanks, 1986] Hanks, Patrick (réd.), *Collins dictionary of the English language*, Collins, London, 1986, 2e édn., 1771 pp.

[Harabagiu et al., 2000] Harabagiu, Sanda, Moldovan, Dan, Psca, Marius, Mihalcea, Rada, Surdeanu, Mihai, Bunescu, Razvan, Gîrju, Roxana, Rus, Vasile et Morarescu, Paul, Falcon *: Boosting Knowledge for Answer Engines*, dans *Proceedings of Text REtrieval Conference*, 2000, pp. 479–488.

[Harman, 1992] Harman, Donna, *Overview of the First Text REtrieval Conference (TREC-1)*, dans *Proceedings of The First Text Retrieval Conference*, TREC-1, 1992, pp. 1–20.

[Hayes, 1977] Hayes, Philip J., *On semantic nets, frames and associations*, dans *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 1977, pp. 99–107.

[Hearst, 1991] Hearst, Marti A., *Noun homograph disambiguation using local context in large corpora*, dans *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and the Text Research*, Oxford, UK, 1991, pp. 1–19.

[Hirst, 1987] HIRST, Grame (réd.), *Semantic interpretation and the resolution of ambiguity*, Studies in Natural Language Processing, Cambridg University Press, Cambridge, Royaume Uni, 1987, 263 pp.

[Hull, 1999] HULL, David A., *Xerox TREC-8 Question Answering Track Report*, dans *Proceedings of The Eighth Text Retrieval Conference*, TREC-8, 1999, pp. 743–752.

[Ide et Véronis, 1998] IDE, Nancy et VÉRONIS, Jean, *Introduction to the special issue on word sense disambiguation : the state of the art*, dans *Computational Linguistics*, tm. 24(1), 1998, pp. 1–40.

[Jacquemin et al., 2002] JACQUEMIN, Bernard, BRUN, Caroline et ROUX, Claude, *Enriching a text by semantic disambiguation for information extraction*, dans DE LOUPY, Claude (réd.), *LREC 2002 Workshop Proceedings. Using Semantics for Information Retrieval and Filtering*, CIRLEP (Centre Interdisciplinaire de Recherches en Linguistique et Psychologie cognitive, Las Palmas de Gran Canaria, Espagne, 2002, pp. –.

[Jacquemin, 1999] JACQUEMIN, Christian, *Syntagmatic and Paradigmatic Representation of Term Variation*, dans *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, ACL'99, 1999, pp. 341–348.

[Justeson et Katz, 1995] JUSTESON, J.S. et KATZ, S.M., *Technical terminology : some linguistic properties and an algorithm for identification in text*, dans *Natural Language Engineering*, tm. 1, 1995, pp. 9–27.

[Kaplan, 1955] KAPLAN, Abraham, *An experimental study of ambiguity and context*, dans *Mechanical Translation*, tm. 2(2), 1955, pp. 39–46.

[Kelly et Stone, 1975] KELLY, Edward F. et STONE, Philip J. (réds.), *Computer recognition of English word senses*, North-Holland, Amsterdam, 1975.

[Kilgarriff, 1994] KILGARRIFF, Adam, *The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary)*, dans *Proceedings of the 6th International Congress of Lexicography*, 1994, pp. 101–106.

[Kim et Moldovan, 1993] KIM, Jun-Tae et MOLDOVAN, Dan I., *PALKA : A System for Lexical Knowledge Acquisition*, dans *Proceedings of the Ninth IEEE Conference an Artificial Intelligence for Applications*, IEEE Computer Society Press, 1993, pp. 171–176.

[Kushmerick et al., 1997] KUSHMERICK, Nicholas, WELD, Daniel S. et DOORENBOS, Robert, *Wrapper Induction for Information Extraction*, dans *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, IJCAI-97, 1997, pp. 729–737.

[Lehnert, 1977] LEHNERT, Wendy, *Human and Computational Question Answering*, dans *Cognitive Science*, tm. 1, 1977, pp. 47–63.

[Lehnert, 1979] LEHNERT, Wendy (réd.), *The process of Question Answering*, Lawrence Erlbaum Associates, 1979.

[Lehnert, 1990] LEHNERT, Wendy, *Symbolic/Subsymbolic Sentence Analysis : Exploiting the Best of Two Worlds*, dans BARNDEN, J. et POLLACK, J. (réds.), *Advances in Connexionist and Natural Computation Theory*, Ablex Publishers, Norwood, NJ, 1990, tm. 1, pp. 135–164.

[Lesk, 1996] LESK, Michael, *Automated Sense Disambiguation Using Machine-readable Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone*, dans *Proceedings of the 1986 SIGDOC Conference*, 1996, pp. 24–26.

[Lewis, 1992] LEWIS, David D., *Text Filtering in MUC-3 and MUC-4*, dans *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, MUC-4, Morgan Kaufmann, San Mateo, McLean, Virginia, États-Unis, 1992, pp. 51–66.

[Liddy et Paik, 1992] LIDDY, Elisabeth E. et PAIK, Woojin, *Statistically-guided word sense disambiguation*, dans *Proceedings of the AAAI Fall Symposium Series*, 1992, pp. 98–107.

[Madhu et Lytle, 1965] MADHU, Swaminathan et LYTLE, Dean W., *A figure of merit technique for the resolution of non-grammatical ambiguity*, dans *Mechanical Translation*, tm. 8(2), 1965, pp. 9–13.

[Makhoul et al., 1999] MAKHOUL, John, KUBALA, Francis, SCHWARTZ, Richard et WEISCHEDEL, Ralph, *Performances measures for information extraction*, dans *Proceedings DARPA Broadcast News Workshop*, LREC 2002, Herndon, Virginia, États-Unis, 1999, pp. 249–252.

[Martinet, 1960] MARTINET, André, *Éléments de linguistique générale*, Colin, Paris, 1960.

[Masterman, 1957] MASTERMAN, Margaret, *The thesaurus in syntax and semantics*, dans *Mechanical Translation*, tm. 4, 1957, pp. 1–2.

[Masterman, 1961] MASTERMAN, Margaret, *Semantic message detection for machine translation using an interlingua*, dans *Proceedings of the International Conference on Machine Translation*, Her Majesty's Stationery Office, London, 1961, pp. 438–475.

[Menon et Modiano, 1993] MENON, Bruno et MODIANO, Nicole, *Eagles. Lexicon Architecture*, Eagles, 1993.

[Meyer et Schvaneveldt, 1975] MEYER, David E. et SCHVANEVELDT, Roger W., *Loci of contextual effects on visual word recognition*, dans RABBITT, P. et DORNIC, S. (réds.), *Attention and performance*, Academic Press, 1975, tm. 5, pp. 98–118.

[Michiels, 1982] MICHIELS, Archibald, *Exploiting a large dictionary data base*, Thèse de doctorat, Université de Liège, Liège, Belgique, 1982.

[Miller, 1985] Miller, George A., *WordNet : a dictionary browser*, dans *Proceedings of the First International Conference on Information in Data*, University of Waterloo, Waterloo, Ontario, États-Unis, 1985.

[Miller et al., 1990] Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek et Miller, Katherine J., *Introduction to WordNet : An on-line lexical database*, dans *International Journal of Lexicography*, tm. 3, 1990, pp. 235–244.

[Moldovan et al., 2000] Moldovan, Dan, Harabagiu, Sanda, Psca, Marius, Mihalcea, Rada, Goodrum, Richard, Gîrju, Roxana et Rus, Vasile, *The structure and performance of an open-domain question answering system*, dans *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, 2000, pp. 563–570.

[Mollá Aliod et al., 2000] Mollá Aliod, D., Schwitter, R., Hess, M. et Fournier, R., *Extrans. An Answer Extraction System*, dans *Traitement Automatique des Langues*, tm. 41(2), 2000, pp. 496–522.

[Panov, 1960] Panov, D, *La traduction mécanique et l'humanité*, dans *Impact*, tm. 10(1), 1960, pp. 17–25.

[Peters, 2002] Peters, Carol (réd.), *Results of the CLEF 2002 Cross-Langage System Evaluation Campaign*, Rome, 2002, 623 pp.

[Pimsleur, 1957] Pimsleur, P., *Semantic frequency counts*, dans *Mechanical Translation*, tm. 4(1-2), 1957, pp. 11–13.

[Pustejovsky, 1991] Pustejovsky, James, *The Generative Lexicon*, dans *Computational Linguistics*, tm. 17, 1991, pp. 409–441.

[Quillian, 1968] Quillian, M. Ross, *Semantic memory*, dans Minsky, M. (réd.), *Semantic information processing*, MIT Press, Cambridge, Massachusetts, 1968, pp. 227–270.

[Reifler, 1955] Reifler, Erwin, *The mechanical determination of meaning*, dans Locke, William N. et Booth, A. Donald (réds.), *Machine translation of languages*, John Wiley & Sons, New York, 1955, pp. 136–164.

[Resnik, 1992] Resnik, Philip, *WordNet and distributional analysis : a class-based approach to statistical discovery*, dans *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, 1992, pp. 48–56.

[Resnik, 1995] Resnik, Philip, *Disambiguating Noun Groupings with Respect to WordNet Senses*, dans *Proceedings of the Third Workshop on Very Large Corpora*, 1995, pp. 54–68.

[Riloff, 1993] Riloff, Ellen, *Automatically Constructing a Dictionary for Information Extracting Tasks*, dans *Proceedings of the Eleventh National Conference on Artificial Intelligence*, AAAI Press / MIT Press, 1993, pp. 811–816.

[Riloff, 1996a] Riloff, Ellen, *Automatically Generating Extraction Patterns from Untagged Text*, dans *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI-96, 1996a, pp. 1044–1049.

[Riloff, 1996b] Riloff, Ellen, *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains*, dans *Artificial Intelligence Journal*, tm. 85(1-2), 1996b, pp. 101–134.

[Riloff et Lorenzen, 1999] Riloff, Ellen et Lorenzen, Jeffrey, *Extraction-based text categorization : generating domain-specific role relationships automatically*, dans Strzalkowski, T. (réd.), *Natural Language Information Retrieval*, Kluwer Academic Publisher, 1999, pp. 167–196, URL citeseer.ist.psu.edu/riloff98extractionbased.html.

[Riloff et Shepherd, 1997] Riloff, Ellen et Shepherd, Jessica, *A Corpus-Based Approach for Building Semantic Lexicons*, dans *Proceedings of the Second Conference on Empirical Methids in Natural Language Processing*, EMNLP-2, 1997, pp. 117–124.

[Roget et Dutch, 1972] Roget, Peter Mark et Dutch, Robert A. (réds.), *Roget's Thesaurus of ENglish Words and Phrases*, Longman, London, 1972, 6^{e} édn.

[Roux, 1999] Roux, Claude, *Phrase-driven parser*, dans *Proceedings of VEXTAL'99*, VEXTAL'99, Venezia, Italia, 1999, pp. 235–240.

[Roux et Jacquemin, 2002] Roux, Claude et Jacquemin, Bernard, *Storing and indexing of each level of the inner structure of a document with binary vector indexes, with the deepest level being the result of natural language processes*, Déposition de proposition de brevet, Xerox, 2002.

[Salton et McFill, 1983] Salton, Gerard et McFill, Michael J. (réds.), *Introduction to Modern Information Retrieval*, Language, Speech and Communication, McGraw-Hill Publishing Company, New York, 1983.

[Schütze, 1992] Schütze, Hinrich, *Dimensions of meaning*, dans *Proceedings of Supercomputing '92*, IEEE Computer Society Press, Los Alomitos, California, 1992, pp. 787–796.

[Segond, 2000] Segond, Frédérique, *Framework and results for French*, dans *Computer and the Humanities. Special Issue on SENSEVAL*, tm. 34(1-2), 2000, pp. 49–60.

[Segond et al., 1997] Segond, Frédérique, Shiller, Anne, Grefenstette, Gregory et Chanod, Jean-Pierre, *An Experiment in Semantic Tagging Using Hidden Markov Model Tagging*, dans *ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications*, 1997.

[Segond et al., 1998] Segond, Frédérique, Aimelet, Élisabeth et Griot, Laurent, *"All you can use !" or how to perform Word Sense Disambiguation with available resources*, dans *Second Workshop on Lexical Semantic System*, 1998.

[Segond et al., 2000] SEGOND, Frédérique, AIMELET, Élisabeth, LUX, Veronika et JEAN, Corinne, *Dictionary-driven Semantic Look-up*, dans *Computer and the Humanities. Special Issue on SENSEVAL*, tm. 34(1-2), 2000, pp. 193–197.

[Sheridan et Ballerini, 1996] SHERIDAN, Páraic et BALLERINI, Jean-Paul, *Experiments in multilingual information retrieval using the Spider system*, dans *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM SIGIR, Zurich, 1996, pp. 58–65.

[Slator, 1992] SLATOR, Brian M., *Senses and Preference*, dans *Computer and Mathematics with Applications*, tm. 23(6/9), 1992, pp. 391–402.

[Snover et al., 2002] SNOVER, Matthew G., JAROSZ, Gaja E. et BRENT, Michael R., *Unsupervised Learning of Morphology Using a Novel Directed Search Algorithm : Taking the First Step*, dans *ACL'02 Workshop Proceedings on Morphological and Phonological Learning*, ACL'02, Philadelphia, PA, USA, 2002.

[Soderland, 1999] SODERLAND, Stephen, *Learning Information Extraction Rules for Semi-Structured and Free Text*, dans *Machine Learning*, tm. 34(1), 1999, pp. 233–272.

[Soderland et al., 1995] SODERLAND, Stephen, FISHER, David, ASELTINE, Jonathan et LEHNERT, Wendy, *CRYSTAL : Inducing a Conceptual Dictionary*, dans *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, IJCAI-95, 1995, pp. 1314–1320.

[Sussna, 1993] SUSSNA, Michael, *Word Sense Disambiguation for Free-text Indesxing using a Massive Semantic Network*, dans *Proceedings of the Second International Conference on Information and Knowledge Bas Management*, 1993, pp. 67–74.

[Tengi, 1998] TENGI, Randee I., *Design and implementation of the WordNet lexical database and searching software*, dans FELLBAUM, Christiane (réd.), *WordNet : an electronic lexical database*, The MIT Press, Cambridge, Massachusetts, 1998, pp. 105–127.

[Trouilleux, 1998] TROUILLEUX, François, *Thingfinder Prototype English Version 2.0*, Rap. tech., Xerox Research Center Europe, 1998.

[Trouilleux, 2001] TROUILLEUX, François, *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*, Thèse de doctorat, Université Clermont 2 Blaise Pascal, Clermont-Ferrand, 2001.

[Viegas et Bouillon, 1994] VIEGAS, Evelyne et BOUILLON, Pierrette, *Semantic Lexicons : The Cornersone for Lexical Choice in NLG*, dans *Proceedings of the 7th International Workshop on Natural Langage Generation*, 1994. Http ://budling.nytud.hu/ kalman/reading/siggen94/siggen94.html.

[Viegas et al., 1999] VIEGAS, Evelyne, MAHESH, Kavi et NIERENBURG, Sergei, *Semantics in Action*, dans SAINT-DIZIER, Patrick (réd.), *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, Kluwer Academic, Dordrecht, 1999, pp. 171–203.

[Voorhees, 1993] VOORHEES, Ellen M., *Using WordNet to disambiguate word senses for text retrieval*, dans *Proceedings of the 16th Annual International ACM SIGIR Conference on Research qnd Development in Information Retrieval*, ACM SIGIR, Pittsburgh, Pennsylvania, 1993, tm. 2, pp. 171–180.

[Voorhees, 1999] VOORHEES, Ellen M., *The TREC-8 Question Answering Track Report*, dans VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceedings of TREC-8*, 1999, pp. 77–82.

[Voorhees, 2000] VOORHEES, Ellen M., *Overview of the TREC-9 Question Answering Track*, dans VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceedings of TREC-9*, 2000, pp. 71–80.

[Voorhees, 2001] VOORHEES, Ellen M., *Overview of the TREC-10 Question Answering Track*, dans VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceedings of TREC-10*, 2001, pp. 42–51.

[Voorhees, 2002] VOORHEES, Ellen M., *Overview of the TREC 2002 Question Answering Track*, dans VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceedings of TREC 2002*, 2002.

[Voorhees et Harman, 1999a] VOORHEES, Ellen M. et HARMAN, Donna, *Overview of the Eighth Text REtrieval Conference (TREC-8)*, dans VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceedings of TREC-8*, 1999a, pp. 1–24.

[Voorhees et Harman, 1999b] VOORHEES, Ellen M. et HARMAN, Donna (réds.), *Proceeding of the Eighth Text REtrieval Conference (TREC-8)*, 1999b.

[Vossen, 1997] VOSSEN, Piek, *EuroWordNet : a multilingual database for information retrieval*, dans *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, CIRLEP (Centre Interdisciplinaire de Recherches en Linguistique et Psychologie cognitive, Zurich, Suisse, 1997.

[Vossen, 1998] VOSSEN, Piek (réd.), *EuroWordNet : a multilingual database with lexical semantic networks*, Kluwer Academic Publishers, New York, 1998. Réédition de *Computer and the Humanities*, 32(2-3), 1998.

[Véronis et Ide, 1990] VÉRONIS, Jean et IDE, Nancy, *Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries*, dans *Proceedings of the 13th International Conference on Computational Linguistics*, COLING'90, Helsinki, Finland, 1990, tm. 2, pp. 389–394.

[Weaver, 1949] WEAVER, Warren, *Translation*, dans LOCKE, William N. et BOOTH, A. Donald (réds.), *Machine translation of languages*, John Wiley and Sons, New York, 1949, pp. 15–23.

[Weiss, 1973] WEISS, Stephen Frederiks, *Learning to disambiguate*, dans *Information Storage and Retrieval*, tm. 9, 1973, pp. 33–41.

[Wilks, 1975] WILKS, Yorick A., *A preferential, pattern-seeking semantics for natural langage inference*, dans *Artificial Intelligence*, tm. 6, 1975, pp. 53–74.

[Wilks et Stevenson, 1996] WILKS, Yorick A. et STEVENSON, Mark, *The grammar of sense : Is word sense tagging much more than part-of-speech tagging ?*, Rap. tech. CS-96-05, MIT, 1996.

[Wilks et al., 1993] WILKS, Yorick A., FASS, Dan, GUO, Cheng-Ming, MACDONALD, James E., PLATE, Tony et SLATOR, Brian A., *Providing Machine Tractable Dictionary Tools*, dans PUSTEJOVSKY, James (réd.), *Semantics and the Lexicon*, Kluwer, London, 1993, pp. 341–401.

[Yarowsky, 1992] YAROWSKY, David, *Word sense disambiguation using statistical models of Roget's categories trained on large corpora*, dans *Proceedings of the 14th International Conference on Computational Linguistics*, COLING, Nantes, France, 1992, pp. 454–460.

[Zock et Mitkov, 1991] ZOCK, M. et MITKOV, R., *How to ask a foreigner questions without knowing his language ? Proposal for a conceptual interface to communicate thought*, dans *Proceedings of Natural Language Processing Pacific Rim Symposium*, Singapour, 1991, pp. 121–130.

Index

Glossaire

actant :
désigne celui qui fait l'action indiquée par le verbe ou le groupe verbal.

actualisation :
réalisation concrète sous la forme d'un mot ou de plusieurs mots d'un sens théorique.

amorçage sémantique :
théorie psycholinguistique selon laquelle l'introduction d'un concept dans un énoncé va influencer et faciliter la compréhension de concepts ultérieurs sémantiquement reliés (*semantic priming*).

amorçage :
technique utilisée pour fournir à un système d'analyse les données dont il a besoin pour fonctionner, et qui consiste à passer par une phase contrôlée d'apprentissage d'informations qui permettront ultérieurement une analyse automatique (*bootstrapping*).

analyseur robuste :
outil d'analyse linguistique automatique qui présente la particularité de tolérer que lui soit présenté un énoncé qui sort de sa compétence sans pour autant provoquer un arrêt du système : simplement, aucun résultat ne sera proposé par l'analyseur.

anaphore :
processus syntaxique consistant à reprendre par un segment (habituellement un pronom ou un adjectif) l'entité actualisée par un autre segment du discours ; nous utilisons couramment le terme anaphore dans le sens plus restreint de *coréférence** anaphorique.

ancrage :
l'ancrage d'un trait est l'entité à laquelle un trait est rattaché – que ce soit un *lexème**, un nœud lexical ou non lexical de l'arbre syntaxique partiel ou une *dépendance syntaxique**.

antonyme :
*lexème** qui possède un sens opposé à un autre par rapport auquel il est envisagé.

***baseline* :**
voir *plancher**.

booléen :
voir *logique booléenne**.

***bootstrapping* :**
voir *amorçage**.

catégorie grammaticale :
classe d'une *unité lexicale** définie sur la base de critères syntaxiques : nom, pronom, verbe, adjectif, déterminant, adverbe, préposition, conjonction, interjection (synonyme : *partie du discours**).

cause :
relation d'*implication** unissant deux verbes dont le premier est causatif et le second résultatif.

***chunk* :**
voir *syntagme minimal**.

cible :
le mot que le système de *désambiguïsation sémantique** doit traiter à un instant donné (*target*).

CIFRE :
Convention Industrielle pour la Formation à la Recherche en Entreprise.

CLEF :
Cross Language Evaluation Forum.

collocation :
association habituelle d'une *unité lexicale** avec une autre dans un énoncé.

correct :
se dit d'une réponse apportée en *gestion de l'information** qui est conforme à la réponse de référence pour la même information demandée.

coréférence :
relation syntaxique unissant deux segments du discours qui dans le contexte désignent la même entité ou font référence à la même entité ; la coréférence *anaphorique** relie un pronom ou un adjectif (qui ne fait référence par lui-même qu'à une entité indéterminée) à l'actualisation de cette entité placée avant dans l'énoncé.

dégradation sur la correspondance :
réduction volontaire des contraintes de correspondance entre l'information contenue dans la question posée au système de *question-réponse** et les réponses proposées.

dégradation sur la question :
réduction volontaire de la quantité d'information extraite de la question par le système de *question-réponse** qui aboutit à une diminution des contraintes informationnelles exigées dans les réponses proposées par le système.

dépendance fonctionnelles :
dépendances que *XIP** extrait pour son fonctionnement interne (c'est-à-dire qui permettent de construire d'autres *dépendances**) ou qui mettent en œuvre un ou plusieurs *mots-outils**.

dépendance syntaxique :
propriété d'ordre syntaxique attribuée par un analyseur syntaxique automatique à un *lexème** ou à la relation qui unit deux ou plusieurs *lexèmes**.

dépendances transitives :
dépendances dont la définition par l'analyseur syntaxique est multiple, mais qui recouvrent une seule relation syntaxico-sémantique malgré une présentation syntaxique différente (par exemple une relation actant-action correspond à la syntaxe du sujet et du complément d'agent).

désambiguïsation concurrente :
cas où l'application de la *désambiguïsation sémantique** permet à plus d'un sens de s'affirmer, c'est-à-dire pour la méthode choisie, lorsque plusieurs règles concurrentes s'appliquent à un même mot dans un même contexte (synonyme : *désambiguïsation incomplète**).

désambiguïsation incomplète :
voir *désambiguïsation concurrente**.

désambiguïsation sémantique :
association d'un *unité lexicale** apparaissant dans un contexte avec sa signification ou sa définition – laquelle peut être distinguée des autres définitions qu'on peut attribuer à ce *lexème** (*word sense disambiguation*).

dictionnaire informatique :
ressource lexicale se présentant sous un format numérique tel que seul un programme informatique (souvent dédié) peut l'exploiter, généralement une base de données optimisées.

dictionnaire sémantique :
ressource lexicale qui privilégie la description des relations sémantiques entre les *unités lexicales** qu'il contient par rapport aux autres informations.

dictionnaire électronique :
version numérisée d'un dictionnaire papier.

DTD :
Document Type Definition, Définition de Type de Document.

enrichissement :
opération effectuée sur un texte ou un énoncé qui consiste à donner à ce texte une présentation différente de son aspect originel tout en lui conservant un sens égal ou similaire ; le résultat de cette opération (synonyme : *expansion**).

EWN :
EuroWordNet.

expansion :
voir *enrichissement**.

expression synonymique :
locution lexicalisée ou non dont la propriété est d'avoir le même sens qu'une *unité lexicale**.

expressions régulières :
formalisme de traitement de chaînes de caractères sous forme de modèles ou patrons très généraux ou très spécifiques, et permettant d'identifier, d'extraire ou de modifier les chaînes de caractères sélectionnées par le patron.

extraction d'information :
domaine de la *gestion de l'information** qui consiste à extraire automatiquement de l'information structurée à partir d'un texte en langage naturel non structuré (*information extraction*).

***F-measure* :**
voir *F-mesure**.

F-mesure :
résultat statistique qui découle de la combinaison de la *précision** et du *rappel**, pondéré par un paramètre β dont la variation à partir de 1.0 détermine si le *rappel** ou la *précision** est de plus de poids (*F-measure*).

fenêtre de réponse :
étendue prévue pour une réponse apportée par un système de *gestion de l'information**, exprimé selon les cas en caractères, en unités lexicales ou en unité linguistiques ou typographiques plus étendues (phrase, paragraphe, texte).

***focus* :**
concept présent dans une question qui englobe l'information attendue en réponse à cette question ; objet de la question.

forme canonique :
forme d'un *lexème** considérée comme non fléchie, et utilisée arbitrairement comme intitulé pour cette *unité lexicale** dans les dictionnaires.

formulaire :
structure hiérarchique d'attributs-valeurs permettant d'exprimer une information et couramment utilisée en *extraction d'information** (*template*).

gestion de l'information :
l'ensemble du domaine du traitement automatique de données appliqué au contenu de textes, et principalement la recherche d'information, le filtrage de textes, l'*extraction d'information** et la tâche de *question-réponse** (*knowledge management*).

groupe syntaxique minimal :
voir *syntagme minimal**.

***hidden Markov model* (HMM)**
voir *modèle de Markov caché**.

HLRT :
Head Left Right Tail : règle qui consiste à identifier les bornes gauche et droite de chaque élément d'information.

HMM :
Hidden Markov Model, voir *Modèle de Markov Caché**.

holonyme :
*lexème** dont le sens, désignant un tout, inclut le sens d'autres termes, considérés comme désignant des parties de ce tout, et appelés *méronymes**.

HTML :
HyperText Markup Language.

hypéronyme :
*lexème** dont le sens, considéré comme plus général, inclut le sens d'autres termes, considérés comme plus spécifiques, et appelés *hyponymes**.

hyponyme :
*lexème** dont le sens, considéré comme plus spécifique, est inclus dans le sens d'un autre terme, considéré comme plus général, et appelé *hypéronyme**.

héritage :
relation sémantique hiérarchique selon laquelle les propriétés sémantiques d'une *unité lexicale** sont héritées par les unités lexicales qui lui sont hiérarchiquement soumises dans une *taxinomie**. Les relations d'héritage sont l'*hyponymie** et l'*hypéronymie**.

IE :
Information Extraction, voir *extraction d'information**.

IFSP :
Incremental Finite-State Parser.

ILI :
Inter-Lingual Index, index inter-langue.

ILPGA :
Institut de Linguistique et de Phonétique Générales et Appliquées.

implication :
relation sémantique unissant un verbe à un autre lorsque la vérité du procès simple dans lequel apparaît le premier provoque obligatoirement la vérité du procès simple dans lequel apparaît le second (par exemple, *il ronfle* implique *il dort*).

incorrect :
se dit d'une réponse apportée en *gestion de l'information** qui n'est pas conforme à la réponse de référence pour la même information demandée.

information extraction :
voir *extraction d'information**.

IR :
Information Retrieval, recherche d'information.

***knowledge management* :**
voir *gestion de l'information**.

LDOCE :
Longman Dictionary of Contemporary English.

lexical :
voir *lexème**.

lexique génératif :
ressource lexicale décrivant les sens du lexique considéré sous la forme de règles de signification relatives [Pustejovsky, 1991].

lexème :
unité de base du lexique constitutive de sens ; le lexème peut être un mot simple, un mot composé ou une locution (synonyme : *unité lexicale**).

lissage :
méthode s'appuyant sur des fréquences de co-occurrences utilisée pour éviter que la probabilité d'apparition d'un sens rare et non représenté dans le corpus soit égale à zéro (*smoothing*).

logique booléenne :
logique binaire s'appliquant à des propositions qui sont donc soit vraies (1) soit fausses (0), fondée sur des opérateurs d'inclusion (ET), de disjonction inclusive ou exclusive (OU) et de négation (NON).

méronyme :
*lexème** dont le sens, désignant une partie d'un tout, est inclus dans le sens d'un autre terme, considérés comme désignant un tout, et appelés *holonyme**.

mesures traditionnelles :
ces mesures sont celles de *précision** et de *rappel**, utilisées pour évaluer la qualité de la plupart des systèmes de *gestion de l'information** depuis ses débuts, à l'exception de la tâche de *question-réponse**.

méthode globale :
pour le système de construction de la *structure informationnelle** et le système d'interrogation développé dans cette thèse, type d'interrogation consistant à interroger la structure informationnelle enrichie au maximum (y compris la *coréférence** *anaphorique** des sujets) en utilisant comme contraintes la présence du *lexical** *focus** et le rejet des réponses pour lesquelles aucune *dépendance** n'est strictement exacte dans la question et la réponse candidate.

Modèle de Markov caché :
modèle statistique de distribution de traits « cachés », tels que des étiquettes de phonèmes ou de *catégories grammaticales**, basé sur des traits observables, tels que des segments acoustiques ou des mots ; les modèles computationnels peuvent être entraînés automatiquement à partir d'un échantillon de données utilisée pour identifier la couche « cachée », basés sur le modèle statistique dérivé des données d'entraînement.

mot-vedette :
intitulé d'un article dans une ressource lexicale, correspondant à la *forme canonique** de l'*unité lexicale** décrite.

mots grammaticaux :
classe de mots fermée qui indiquent des relations grammaticales entre syntagme ou phrases, ou qui indiquent la frontière d'un syntagme ; elle regroupe les déterminants, les auxiliaires et copules, les conjonctions et les prépositions (synonymes : *mots vides**, *mots-outils**).

mots vides :
voir *mots grammaticaux**.

mots-outils :
voir *mots grammaticaux**.

MUC :
Message Understanding Conference.

***multi-slot structure* :**
voir *structure d'information combinée**.

***NLP* :**
Natural Language Processing, voir *TAL**.

NTCIR :
NII-NACSIS Test Collection for IR systems.

OHFD :
Oxford Hachette French Dictionary.

***overgeneration* :**
voir *surgénération**.

parasynonyme :
il s'agit dans le *Dictionnaire des verbes et des mots* de [Dubois et Dubois-Charlier, 1997], *unité* ou expression *lexicale** sémantiquement reliée au *mot-vedette** soit comme *synonyme**, soit comme *holonyme** ou *hypéronyme**, soit comme *synonyme** d'un *holonyme** ou d'un *hypéronyme**.

partie du discours :
voir *catégorie grammaticale** (*part-of-speech*).

***part-of-speech* :**
voir *catégorie grammaticale**.

partonomie :
structure hiérarchique sémantique qui met en œuvre les relations d'*holonymie** et de *méronymie** et classe les *lexèmes** selon ces relations (voir *taxinomie**).

patient :
désigne celui qui subit l'action indiquée par le verbe ou le groupe verbal.

plancher :
référence permettant de juger du niveau d'efficacité d'un système, et consistant ici à questionner la base documentaire sans enrichissement à l'aide des mots pleins des questions utilisés comme des mots-clefs (*baseline*).

précision :
rapport du nombre des réponses *correctes** obtenues au nombre total des réponses (*correctes** et *incorrectes**) obtenues.

présupposition :
relation d'*implication** entre deux verbes dont l'application du procès du premier implique la réalisation préalable du procès du second

QA :
Question Answering, voir *question-réponse**.

***question answering* :**
voir *question-réponse**.

question-réponse :
ce domaine de la *gestion de l'information** s'intéresse à l'interrogation en langage naturel d'une base documentaire. Son but est de permettre de poser des questions comme à un être humain au système qui s'efforce d'y apporter soit un court fragment de texte contenant la réponse à la question posée, soit cette réponse uniquement (*question answering*).

racinisation :
simplification automatique d'une *unité lexicale** à sa racine ou à un noyau de cette *unité** considéré comme sa racine pour permettre de trouver dans un texte toutes les *unités lexicales** possédant une même racine, qui seront dès lors considérées comme appartenant au même champ sémantique (*stemming*).

rappel :
rapport des réponses *correctes** obtenues au nombre des réponses *correctes** réellement présentes dans le texte interrogé (*recall*).

***recall* :**
voir *rappel**.

règle sémantique :
règle créée pour la *désambiguïsation sémantique**, et dans laquelle les contraintes sur le contexte lexical ne porte que sur des traits sémantiques de ce contexte (par opposition aux règles lexicales et aux règles de *sous-catégorisation**).

scalarité :
relation sémantique unissant des adjectifs appartenant à une échelle de gradation de sens dont les extrémités ont des sens opposés.

score :
résultat évaluatif donné à la réponse à une requête apportée par un système de *question-réponse**, et correspondant au rapport à 1 du rang de la première bonne réponse parmi les cinq premières fournies, ou 0 ; moyenne des scores pour chaque question lorsque l'évaluation est menée sur un ensemble de questions.

segmentation :
découpage d'un texte en *unités lexicales** (*tokenization*).

***semantic priming* :**
voir *amorçage sémantique**.

semi-auxiliaire :
verbes qui, construits avec un infinitif, parfois avec un participe ou un gérondif, perdent plus ou moins leur signification propre et servent à exprimer diverses nuances de temps, d'aspect ou d'autres modalités de l'action.

***single-slot structure* :**
voir *structure d'extraction simple**.

***smoothing* :**
voir *lissage**.

sous-catégorisation :
indication limitative sur le schéma syntaxique (ou parfois syntaxico-sémantique) propre à une *unité lexicale** (par exemple *transitif* pour un verbe donné) ; lorsque cette indication est propre à une acception de l'*unité lexicale**, l'apparition de ce schéma syntaxique dans un énoncé permet de désigner le sens de l'*unité lexicale** dans cet énoncé.

***stemming* :**
voir *racinisation**.

structure d'extraction combinée :
formulaire d'extraction d'une information complexe permettant de mettre en rapport différents éléments de même nature d'une même information (*multi-slot structure*).

structure d'extraction simple :
formulaire d'extraction d'une information complexe incapable de mettre en rapport différents éléments de même nature d'une même information, et de ce fait limité à un seul élément informationnel de même nature (*single-slot structure*).

structure informationnelle :
agencement constitué à partir d'une base documentaire textuelle par l'identification et le classement de chaque information qu'elle contient ainsi que par la détermination des liens qui unissent les différentes informations. Dans le cas présent, les *enrichissements** font partie de cet agencement d'informations ; ils ont été effectués pour faciliter l'accès aux éléments informationnels.

structure plate :
structure dans laquelle les éléments ne sont pas classés les uns par rapport aux autres, ni hiérarchisés, mais constituent une simple liste.

suffixation :
constitution d'une *unité lexicale** dérivée d'une autre par l'obtention automatique de sa racine et la concaténation de cette racine avec un suffixe.

surgénération :
se dit lorsque la réponse apportée par un système de *gestion de l'information** contient une ou plusieurs données surnuméraires par rapport à la réponse de référence de la même information demandée (*overgeneration*).

synonyme :
nous considérons deux *lexèmes** comme synonymes si un de leur sens au moins est commun de manière à permettre une interchangeabilité dans un énoncé sans modification du sens, même si un remaniement syntaxique est nécessaire.

synonymie aveugle :
enrichissement synonymique pratiqué sur la base documentaire après son analyse morpho-syntaxique, mais sans *désambiguïsation sémantique**, et donc où les *synonymes** ne sont pas distribués en fonction du sens des *cibles** dans les énoncés.

synonymie simple :
relation sémantique de *synonymie** unissant des *unités lexicales** constituant l'entrée d'une ressource lexicale, à l'exclusion de locutions ou expressions à mots multiples.

***synset* :**
ensemble d'*unités lexicales** considérées par *WordNet* ou *EuroWordNet* comme *synonymes** pour un de leurs sens au moins.

syntagme minimal :
correspond à un groupe de la grammaire traditionnelle (groupe nominal, verbal ou prépositionnel), à cette différence près qu'il est tronqué de sa partie droite au-delà de la tête du groupe (*chunk*).

TAL :
Traitement Automatique du Langage naturel (*NLP*).

target
voir *cible**.

taxinomie :
classification hiérarchique sémantique des entrées d'une ressource lexicale qui met en œuvre les relations sémantiques d'*héritage**, c'est-à-dire la *méronymie** et l'*holonymie**.

***template* :**
voir *formulaire**.

texte libre :
texte en langage écrit selon les règles grammaticales en vigueur dans la langue utilisée mais sans autre contrainte.

texte semi-structuré :
texte en langage écrit qui suit rarement la grammaire de la langue et se présente sous un style plus ou moins télégraphique, sans règle rigide ni sans forme prédéfinie ; il est en général porteur d'une information utilitaire.

texte structuré :
texte dans un langage écrit qui répond à des règles très strictes ; l'information dont il est porteur est régulière dans sa nature, sa présence, sa position.

thesaurus :
ressource lexicale monolingue traitant exclusivement et systématiquement la relation de *synonymie** entre les *unités lexicales**.

***tokenization* :**
voir *segmentation**.

traitement homonymique :
mode de présentation d'une ressource lexicale consistant à traiter chacun des sens des lemmes polysémiques sous autant d'entrées, c'est-à-dire à considérer les différents sens d'un même *lexème** comme un ensemble d'homonymes.

transducteur :
dispositif algorithmique qui représente un ensemble de séquences en entrée et qui leur associe des séquences produites en sortie.

TREC :
Text REtrieval Conference.

troponymie :
relation d'*implication** qui relie deux verbes dont l'un décrit une réalisation particulière du procès de l'autre.

unité lexicale :
voir *lexème**.

WN :
WordNet.

***word sense disambiguation* :**
voir *désambiguïsation sémantique**.

***wrapper* :**
procédure logicielle spécifique à un type de structure de ressource informationnelle, et qui traduit la réponse à une requête donnée en une nouvelle *structure d'information simple** ou *combinée** selon la structure de base du document et le sujet de l'information sélectionné.

WSD :
Word Sense Disambiguation, voir *désambiguïsation sémantique**.

XIP :
Xerox Incremental Parser.
XML :
eXtensible Markup Language.
XRCE :
Xerox Research Centre Europe.

Annexes

Annexe A

Méthode de stockage de l'information

Pour mettre en œuvre notre méthode de gestion de l'information, un système d'indexation est nécessaire. Il permet non seulement de stocker et de classifier les données identifiées lors de l'analyse des documents ainsi que celles issues de la phase d'enrichissement, mais aussi de retrouver les fragments de texte qui contiennent l'information recherchée lors de l'interrogation.

La technique d'indexation que nous exploitons [Roux et Jacquemin, 2002] propose le stockage de l'information sous forme de chaînes de bits reliées au texte ou au fragment de texte qui contient cette information. Chacun des bits de la chaîne correspond à un élément d'information : unité lexicale, dépendance, trait, etc.

Par ailleurs, cette technique présente la particularité de constituer une indexation sur plusieurs plans, c'est-à-dire que chaque niveau de découpage du document par l'analyseur correspond à un niveau d'indexation du système. Le niveau le plus général sera donc le document lui-même, puis le paragraphe, la phrase, la dépendance et enfin l'unité lexicale. Chaque représentant d'un de ces niveaux est donc à l'origine d'une chaîne de bits correspondant à l'information qu'il contient.

La figure A.1 page suivante montre de quelle manière chaque élément informatif est conservé par le système. L'indexation au niveau du mot est effectuée par une chaîne de bits dont seul celui qui correspond au lemme de ce mot est activé. Il est toutefois possible de trouver d'autres bits activés : traits portant sur ce mot, enrichissements synonymiques, etc. Cette chaîne de bits pointe vers la position du mot dans le document (*offset*), et permet également de connaître la position relative du mot dans la phrase.

L'indexation au niveau de la dépendance donne naissance à une nouvelle chaîne de bits qui permet un accès à la localisation de phrase[1] du document dans laquelle se trouve la dépendance. Chaque élément informationnel de la dépendance active le bit correspondant. Les arguments de la dépendance sont ordonnés en fonction de leur position relative dans la phrase, indiquée au niveau du mot. Le niveau de la dépendance, comme les niveaux supérieurs, est une structure plate[2] qui permet toutefois d'accéder au niveau inférieur.

```
[##]> echo "Édouard détruisit la flotte française." | xip
```

Dépendances extraites par XIP :

```
SUBJ(<détruisit^détruire :1>,<Édouard^Édouard :0>)
VARG_DIR(<détruisit^détruire :1>,<flotte^flotte :3>)
NMOD_ADJ(<flotte^flotte :3>,<française^français :4>)
```

Indexation au niveau du mot :

Réalité	Représentation	Position
Édouard	...00000001...	0
détruire	...00001000...	1
flotte	...00000100...	3
français	...00000010...	4

Indexation au niveau de la dépendance :

Réalité	Représentation	Position
Édouard	...00000001...	0
détruire	...00001000...	1
SUBJ	...01000000...	
SUBJ(détruire,Édouard)	...01001001...	0-1

FIG. A.1 – Indexation multiplan : exemple.

Les niveaux d'indexation au niveau du paragraphe et du document procèdent de la même manière. Ils pointent respectivement vers la position du paragraphe et vers le document complet, et permettent d'avoir accès aux ni-

[1] La localisation d'une dépendance n'a pas de sens car les lexèmes qui en constituent les arguments ne sont pas forcément contigus. De ce fait, la localisation de la phrase dans laquelle apparaît cette dépendance est la solution la plus pratique.

[2] C'est-à-dire que l'information est donnée sans être structurée ni catégorisée. Seul un accès au niveau du mot permet de structurer les éléments d'information les uns par rapport aux autres. Dans l'exemple, le vecteur correspondant à la dépendance ne permet de connaître l'ordre des arguments que grâce à la position relative des arguments donnée au niveau du mot.

veaux d'indexation inférieurs lorsqu'une structure plus détaillée que la structure plate est nécessaire.

Cette technique d'indexation s'adapte bien aux besoins de notre tâche. En effet, elle est conçue comme un moyen d'accéder rapidement à l'information d'un ou plusieurs documents par simple comparaison des chaînes de bits correspondant à l'information recherchée et à l'information des documents stockée au niveau désiré. Par ailleurs, les données issues de l'analyse et de l'enrichissement sont facilement ajoutées à l'index par simple adjonction du bit correspondant dans les vecteurs de bits du texte analysé à chaque niveau : lexème, trait, expression sous la forme d'une dépendance, etc. La présence de l'enrichissement à l'intérieur même de l'index permet de conserver le texte original et d'y avoir accès au travers de ses enrichissements.

En définitive, on obtient donc une hiérarchie de vecteurs informationnels pointant vers la base documentaire, le premier décrivant l'ensemble de cette base et renvoyant aux documents, chacun représenté par un numéro d'index. Depuis le vecteur d'un document, on a accès au contenu de ses paragraphes au travers de leur index respectif, puis aux phrases par une même démarche, aux dépendances et enfin aux unités lexicales elles-mêmes.

Annexe B

Typologie des questions de TREC-8

who	sujet	d'un copule	25
		d'un autre verbe	22
what	sujet	d'un copule	35
		d'un autre verbe	1
	objet		5
	épithète		19
	adjectif avec préposition		4
	pronom avec préposition		1
which	avec préposition		1
	sans préposition		9
whom			1
how			1
how	much		4
	many		18
how	+ adjectif		8
when			19
where			21
why			2
Pas d'interrogation			4

Annexe C

Ensemble des résultats de l'interrogation

C.1 Évaluation de type question-réponse

Type	Score	Pas de réponse
Avec coréférence - Focus absent - Sans rejet des réponses sans concordances		
Syntaxe	424	113
Synonymes	445	109
D	462	105
D-Syn	464	105
D-B	467	104
D-EWN-M	487	100
D-Dér	467	104
D-B-EWN-M-Dér	502	97
Tous types	504	97
Avec coréférence - Focus absent - Rejet des réponses au seuil de 0		
Syntaxe	414	115
Synonymes	430	112
D	447	108
D-Syn	449	108
D-B	452	107
D-EWN-M	472	103
D-Dér	452	107
D-B-EWN-M-Dér	487	100
Tous types	489	100

Type	Score	Pas de réponse
Avec coréférence - Focus absent - Rejet des réponses au seuil de 10		
Syntaxe	409	116
Synonymes	425	113
D	442	109
D-Syn	444	109
D-B	447	108
D-EWN-M	467	104
D-Dér	447	108
D-B-EWN-M-Dér	482	101
Tous types	484	101
Avec coréférence - Focus absent - Rejet des réponses au seuil de 20		
Syntaxe	380	123
Synonymes	394	120
D	408	117
D-Syn	408	117
D-B	413	116
D-EWN-M	433	112
D-Dér	413	116
D-B-EWN-M-Dér	448	109
Tous types	448	109
Avec coréférence - Focus absent - Rejet des réponses au seuil de 30		
Syntaxe	370	125
Synonymes	379	123
D	393	120
D-Syn	393	120
D-B	398	119
D-EWN-M	418	115
D-Dér	398	119
D-B-EWN-M-Dér	433	112
Tous types	433	112
Avec coréférence - Focus absent - Rejet des réponses au seuil de 40		
Syntaxe	350	129
Synonymes	359	127
D	373	124
D-Syn	373	124
D-B	378	123
D-EWN-M	393	120
D-Dér	378	123
D-B-EWN-M-Dér	403	118
Tous types	403	118

Type	Score	Pas de réponse
Avec coréférence - Focus présent - Sans rejet des réponses sans concordances		
Syntaxe	365	126
Synonymes	379	123
D	398	119
D-Syn	398	119
D-B	403	118
D-EWN-M	423	114
D-Dér	403	118
D-B-EWN-M-Dér	438	111
Tous types	438	111
Avec coréférence - Focus présent - Rejet des réponses au seuil de 0		
Syntaxe	360	127
Synonymes	369	125
D	388	121
D-Syn	388	121
D-B	393	120
D-EWN-M	413	116
D-Dér	393	120
D-B-EWN-M-Dér	428	113
Tous types	428	113
Avec coréférence - Focus présent - Rejet des réponses au seuil de 10		
Syntaxe	360	127
Synonymes	369	125
D	388	121
D-Syn	388	121
D-B	393	120
D-EWN-M	413	116
D-Dér	393	120
D-B-EWN-M-Dér	428	113
Tous types	428	113
Avec coréférence - Focus présent - Rejet des réponses au seuil de 20		
Syntaxe	355	128
Synonymes	364	126
D	378	123
D-Syn	378	123
D-B	383	122
D-EWN-M	403	118
D-Dér	383	122
D-B-EWN-M-Dér	418	115
Tous types	418	115

Type	Score	Pas de réponse
Avec coréférence - Focus présent - Rejet des réponses au seuil de 30		
Syntaxe	355	128
Synonymes	364	126
D	378	123
D-Syn	378	123
D-B	383	122
D-EWN-M	403	118
D-Dér	383	122
D-B-EWN-M-Dér	418	115
Tous types	418	115
Avec coréférence - Focus présent - Rejet des réponses au seuil de 40		
Syntaxe	335	132
Synonymes	344	130
D	358	127
D-Syn	358	127
D-B	363	126
D-EWN-M	378	123
D-Dér	363	126
D-B-EWN-M-Dér	388	121
Tous types	388	121
Référence	295	139
Synonymes aveugles	303	137
Sans coréférence - Focus absent - Sans rejet des réponses sans concordances		
Syntaxe	334	131
Synonymes	350	128
D	368	124
D-Syn	370	124
D-B	367	124
D-EWN-M	388	120
D-Dér	368	124
D-B-EWN-M-Dér	392	119
Tous types	394	119
Sans coréférence - Focus absent - Rejet des réponses au seuil de 0		
Syntaxe	324	133
Synonymes	340	130
D	358	126
D-Syn	360	126
D-B	357	126
D-EWN-M	378	122
D-Dér	358	126
D-B-EWN-M-Dér	382	121
Tous types	384	121

Type	Score	Pas de réponse
Sans coréférence - Focus absent - Rejet des réponses au seuil de 10		
Syntaxe	319	134
Synonymes	335	131
D	353	127
D-Syn	355	127
D-B	352	127
D-EWN-M	373	123
D-Dér	353	127
D-B-EWN-M-Dér	377	122
Tous types	379	122
Sans coréférence - Focus absent - Rejet des réponses au seuil de 20		
Syntaxe	295	140
Synonymes	309	137
D	324	134
D-Syn	324	134
D-B	323	134
D-EWN-M	344	130
D-Dér	324	134
D-B-EWN-M-Dér	348	129
Tous types	348	129
Sans coréférence - Focus absent - Rejet des réponses au seuil de 30		
Syntaxe	285	142
Synonymes	294	140
D	309	137
D-Syn	309	137
D-B	308	137
D-EWN-M	329	133
D-Dér	309	137
D-B-EWN-M-Dér	333	132
Tous types	333	132
Sans coréférence - Focus absent - Rejet des réponses au seuil de 40		
Syntaxe	270	145
Synonymes	279	143
D	294	140
D-Syn	294	140
D-B	294	140
D-EWN-M	309	137
D-Dér	294	140
D-B-EWN-M-Dér	309	137
Tous types	309	137

Type	Score	Pas de réponse
Sans coréférence - Focus présent - Sans rejet des réponses sans concordances		
Syntaxe	280	143
Synonymes	289	141
D	309	137
D-Syn	309	137
D-B	308	137
D-EWN-M	329	133
D-Dér	309	137
D-B-EWN-M-Dér	333	132
Tous types	333	132
Sans coréférence - Focus présent - Rejet des réponses au seuil de 0		
Syntaxe	275	144
Synonymes	284	142
D	304	138
D-Syn	304	138
D-B	303	138
D-EWN-M	324	134
D-Dér	304	138
D-B-EWN-M-Dér	328	133
Tous types	328	133
Sans coréférence - Focus présent - Rejet des réponses au seuil de 10		
Syntaxe	275	144
Synonymes	284	142
D	304	138
D-Syn	304	138
D-B	303	138
D-EWN-M	324	134
D-Dér	304	138
D-B-EWN-M-Dér	328	133
Tous types	328	133
Sans coréférence - Focus présent - Rejet des réponses au seuil de 20		
Syntaxe	270	145
Synonymes	279	143
D	294	140
D-Syn	294	140
D-B	293	140
D-EWN-M	314	136
D-Dér	294	140
D-B-EWN-M-Dér	318	135
Tous types	318	135

Type	Score	Pas de réponse
Sans coréférence - Focus présent - Rejet des réponses au seuil de 30		
Syntaxe	270	145
Synonymes	279	143
D	294	140
D-Syn	294	140
D-B	293	140
D-EWN-M	314	136
D-Dér	294	140
D-B-EWN-M-Dér	318	135
Tous types	318	135
Sans coréférence - Focus présent - Rejet des réponses au seuil de 40		
Syntaxe	255	148
Synonymes	264	146
D	279	143
D-Syn	279	143
D-B	279	143
D-EWN-M	294	140
D-Dér	279	143
D-B-EWN-M-Dér	294	140
Tous types	294	140

C.2 Évaluation de type extraction d'information

Type	**Précision**	**Rappel**	**F-m1**	**F-m2**	**F-m3**	**Exact**	**1 exact**	**Faux**	**Sans**	**Rejet**	**Rej. exact**
Avec coréférence - Focus absent - Sans rejet des réponses sans concordances											
Référence	83.18%	35.74%	65.73	50.00	40.34	89	78	18	116	0	0
Synonymes aveugles	81.20%	38.15%	66.25	51.91	42.68	95	81	22	113	0	0
Syntaxe	31.63%	42.17%	33.29	36.14	39.53	105	82	227	110	0	0
Synonymes	30.41%	44.58%	32.48	36.16	40.78	111	86	254	106	0	0
D	31.51%	46.18%	33.65	37.46	42.25	115	90	250	102	0	0
D-Syn	30.34%	46.18%	32.58	36.62	41.82	115	90	264	102	0	0
D-B	31.45%	46.99%	33.68	37.68	42.76	117	91	255	100	0	0
D-EWN-M	32.35%	48.19%	34.62	38.71	43.89	120	95	251	97	0	0
D-Dér	31.28%	46.99%	33.52	37.56	42.70	117	91	257	101	0	0
D-B-EWN-M-Dér	32.04%	49.80%	34.50	38.99	44.83	124	97	263	93	0	0
Tous enrichissements	31.00%	49.80%	33.53	38.21	44.41	124	97	276	93	0	0
Avec coréférence - Focus absent - Rejet des réponses au seuil de 0											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	32.89%	40.16%	34.13	36.17	38.46	100	79	204	113	28	8
Synonymes	31.33%	41.77%	32.97	35.80	39.16	104	82	228	110	33	9
D	32.63%	43.37%	34.33	37.24	40.69	108	86	223	106	34	10
D-Syn	31.40%	43.37%	33.23	36.42	40.30	108	86	236	106	35	10
D-B	32.64%	44.18%	34.44	37.54	41.26	110	87	227	104	35	10
D-EWN-M	33.53%	45.38%	35.38	38.57	42.39	113	91	224	101	34	10
D-Dér	32.45%	44.18%	34.27	37.41	41.20	110	87	229	105	35	10
D-B-EWN-M-Dér	33.33%	46.99%	35.39	39.00	43.43	117	93	234	97	36	10
Tous enrichissements	32.23%	46.99%	34.39	38.24	43.05	117	93	246	97	37	10

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Avec coréférence - Focus absent - Rejet des réponses au seuil de 10											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	33.00%	40.16%	34.22	36.23	38.49	100	79	203	114	29	8
Synonymes	31.42%	41.77%	33.06	35.86	39.19	104	82	227	111	34	9
D	32.73%	43.37%	34.42	37.31	40.72	108	86	222	107	35	10
D-Syn	31.49%	43.37%	33.31	36.49	40.33	108	86	235	107	36	10
D-B	32.74%	44.18%	34.53	37.61	41.29	110	87	226	105	36	10
D-EWN-M	33.63%	45.38%	35.47	38.63	42.42	113	91	223	102	35	10
D-Dér	32.54%	44.18%	34.35	37.48	41.23	110	87	228	106	36	10
D-B-EWN-M-Dér	33.43%	46.99%	35.48	39.07	43.46	117	93	233	98	37	10
Tous enrichissements	32.32%	46.99%	34.47	38.30	43.08	117	93	245	98	38	10
Avec coréférence - Focus absent - Rejet des réponses au seuil de 20											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	84.00%	33.73%	64.71	48.14	38.32	84	74	16	120	232	20
Synonymes	82.41%	35.74%	65.35	49.86	40.31	89	77	19	117	257	22
D	80.53%	36.55%	64.91	50.28	41.03	91	80	22	114	252	22
D-Syn	79.31%	36.95%	64.52	50.41	41.37	92	80	24	114	263	23
D-B	77.50%	37.35%	63.79	50.41	41.67	93	81	27	111	252	22
D-EWN-M	80.67%	38.55%	66.21	52.17	43.05	96	85	23	109	252	22
D-Dér	80.87%	37.35%	65.59	51.10	41.85	93	81	22	113	259	23
D-B-EWN-M-Dér	78.74%	40.16%	66.05	53.19	44.52	100	87	27	105	260	23
Tous enrichissements	78.29%	40.56%	66.01	53.44	44.89	101	87	28	105	271	24

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Avec coréférence - Focus absent - Rejet des réponses au seuil de 30											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	85.11%	32.13%	64.00	46.65	36.70	80	72	14	122	238	22
Synonymes	84.85%	33.73%	65.12	48.28	38.36	84	74	15	120	266	25
D	81.90%	34.54%	64.28	48.59	39.06	86	77	19	117	260	24
D-Syn	82.08%	34.94%	64.64	49.01	39.47	87	77	19	117	273	26
D-B	78.57%	35.34%	63.13	48.75	39.71	88	78	24	114	260	24
D-EWN-M	81.98%	36.55%	65.66	50.56	41.10	91	82	20	112	260	24
D-Dér	82.24%	35.34%	64.99	49.44	39.89	88	78	19	116	267	25
D-B-EWN-M-Dér	79.83%	38.15%	65.52	51.63	42.60	95	84	24	108	268	25
Tous enrichissements	80.00%	38.55%	65.84	52.03	43.01	96	84	24	108	280	26
Avec coréférence - Focus absent - Rejet des réponses au seuil de 40											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	86.36%	30.52%	63.23	45.10	35.06	76	68	12	126	244	22
Synonymes	86.02%	32.13%	64.41	46.78	36.73	80	70	13	124	272	24
D	82.83%	32.93%	63.57	47.13	37.44	82	73	17	120	266	23
D-Syn	83.00%	33.33%	63.94	47.56	37.86	83	73	17	120	279	25
D-B	79.81%	33.33%	62.41	47.03	37.73	83	74	21	118	268	24
D-EWN-M	82.69%	34.54%	64.66	48.73	39.09	86	77	18	116	267	23
D-Dér	83.17%	33.73%	64.32	48.00	38.29	84	74	17	119	273	24
D-B-EWN-M-Dér	81.65%	35.74%	64.96	49.72	40.27	89	79	20	114	278	25
Tous enrichissements	81.82%	36.14%	65.31	50.14	40.69	90	79	20	114	290	26

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Avec coréférence - Focus présent - Sans rejet des réponses sans concordances											
Référence	83.18%	35.74%	65.73	50.00	40.34	89	78	18	116	0	0
Synonymes aveugles	81.20%	38.15%	66.25	51.91	42.68	95	81	22	113	0	0
Syntaxe	83.51%	32.53%	63.58	46.82	37.05	81	72	16	122	0	0
Synonymes	81.31%	34.94%	64.25	48.88	39.44	87	75	20	119	0	0
D	78.26%	36.14%	63.47	49.45	40.50	90	79	25	115	0	0
D-Syn	77.78%	36.55%	63.46	49.73	40.88	91	79	26	115	0	0
D-B	74.80%	36.95%	62.08	49.46	41.11	92	80	31	112	0	0
D-EWN-M	78.51%	38.15%	64.80	51.35	42.52	95	84	26	110	0	0
D-Dér	78.63%	36.95%	64.16	50.27	41.33	92	80	25	114	0	0
D-B-EWN-M-Dér	76.15%	39.76%	64.37	52.24	43.96	99	86	31	106	0	0
Tous enrichissements	75.76%	40.16%	64.35	52.49	44.33	100	86	32	106	0	0
Avec coréférence - Focus présent - Rejet des réponses au seuil de 0											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	84.78%	31.33%	63.21	45.75	35.85	78	70	14	124	5	1
Synonymes	83.67%	32.93%	63.96	47.26	37.48	82	72	16	122	9	2
D	81.73%	34.14%	63.91	48.16	38.64	85	76	19	118	11	3
D-Syn	81.13%	34.54%	63.89	48.45	39.02	86	76	20	118	11	3
D-B	78.38%	34.94%	62.77	48.33	39.30	87	77	24	115	12	3
D-EWN-M	81.82%	36.14%	65.31	50.14	40.69	90	81	20	113	11	3
D-Dér	82.08%	34.94%	64.64	49.01	39.47	87	77	19	117	11	3
D-B-EWN-M-Dér	79.66%	37.75%	65.19	51.23	42.19	94	83	24	109	12	3
Tous enrichissements	79.83%	38.15%	65.52	51.63	42.60	95	83	24	109	13	3

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Avec coréférence - Focus présent - Rejet des réponses au seuil de 10											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	84.78%	31.33%	63.21	45.75	35.85	78	70	14	124	5	1
Synonymes	83.67%	32.93%	63.96	47.26	37.48	82	72	16	122	9	2
D	81.73%	34.14%	63.91	48.16	38.64	85	76	19	118	11	3
D-Syn	81.13%	34.54%	63.89	48.45	39.02	86	76	20	118	11	3
D-B	78.38%	34.94%	62.77	48.33	39.30	87	77	24	115	12	3
D-EWN-M	81.82%	36.14%	65.31	50.14	40.69	90	81	20	113	11	3
D-Dér	82.08%	34.94%	64.64	49.01	39.47	87	77	19	117	11	3
D-B-EWN-M-Dér	79.66%	37.75%	65.19	51.23	42.19	94	83	24	109	12	3
Tous enrichissements	79.83%	38.15%	65.52	51.63	42.60	95	83	24	109	13	3
Avec coréférence - Focus présent - Rejet des réponses au seuil de 20											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	84.62%	30.92%	62.81	45.29	35.42	77	69	14	125	6	1
Synonymes	83.51%	32.53%	63.58	46.82	37.05	81	71	16	123	10	2
D	81.37%	33.33%	63.17	47.29	37.80	83	74	19	120	13	3
D-Syn	80.77%	33.73%	63.16	47.59	38.18	84	74	20	120	13	3
D-B	77.98%	34.14%	62.04	47.49	38.46	85	75	24	117	14	3
D-EWN-M	81.48%	35.34%	64.61	49.30	39.86	88	79	20	115	13	3
D-Dér	81.73%	34.14%	63.91	48.16	38.64	85	75	19	119	13	3
D-B-EWN-M-Dér	79.31%	36.95%	64.52	50.41	41.37	92	81	24	111	14	3
Tous enrichissements	79.49%	37.35%	64.85	50.82	41.78	93	81	24	111	15	3

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Avec coréférence - Focus présent - Rejet des réponses au seuil de 30											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	84.62%	30.92%	62.81	45.29	35.42	77	69	14	125	6	1
Synonymes	84.38%	32.53%	63.98	46.96	37.09	81	71	15	123	11	3
D	81.37%	33.33%	63.17	47.29	37.80	83	74	19	120	13	3
D-Syn	81.55%	33.73%	63.54	47.73	38.22	84	74	19	120	14	4
D-B	77.98%	34.14%	62.04	47.49	38.46	85	75	24	117	14	3
D-EWN-M	81.48%	35.34%	64.61	49.30	39.86	88	79	20	115	13	3
D-Dér	81.73%	34.14%	63.91	48.16	38.64	85	75	19	119	13	3
D-B-EWN-M-Dér	79.31%	36.95%	64.52	50.41	41.37	92	81	24	111	14	3
Tous enrichissements	79.49%	37.35%	64.85	50.82	41.78	93	81	24	111	15	3
Avec coréférence - Focus présent - Rejet des réponses au seuil de 40											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	107	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	117	0
Syntaxe	85.88%	29.32%	61.97	43.71	33.77	73	65	12	129	12	1
Synonymes	85.56%	30.92%	63.22	45.43	35.45	77	67	13	127	17	2
D	82.29%	31.73%	62.40	45.80	36.17	79	70	17	123	19	2
D-Syn	82.47%	32.13%	62.79	46.24	36.60	80	70	17	123	20	3
D-B	79.21%	32.13%	61.26	45.71	36.46	80	71	21	121	22	3
D-EWN-M	82.18%	33.33%	63.55	47.43	37.83	83	74	18	119	20	2
D-Dér	82.65%	32.53%	63.18	46.69	37.02	81	71	17	122	19	2
D-B-EWN-M-Dér	81.13%	34.54%	63.89	48.45	39.02	86	76	20	117	24	3
Tous enrichissements	81.31%	34.94%	64.25	48.88	39.44	87	76	20	117	25	3

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus absent - Sans rejet des réponses sans concordances											
Référence	83.75%	26.91%	58.88	40.73	31.13	67	59	13	136	0	0
Synonymes aveugles	80.90%	28.92%	59.50	42.60	33.18	72	61	17	133	0	0
Syntaxe	31.75%	32.13%	31.82	31.94	32.05	80	64	172	125	0	0
Synonymes	29.68%	33.73%	30.41	31.58	32.84	84	67	199	121	0	0
D	30.74%	34.94%	31.50	32.71	34.01	87	70	196	117	0	0
D-Syn	29.29%	34.94%	30.27	31.87	33.64	87	70	210	116	0	0
D-B	30.45%	35.34%	31.32	32.71	34.24	88	71	201	115	0	0
D-EWN-M	31.49%	36.55%	32.38	33.83	35.41	91	74	198	112	0	0
D-Dér	30.34%	35.34%	31.23	32.65	34.21	88	70	202	117	0	0
D-B-EWN-M-Dér	30.79%	37.35%	31.91	33.76	35.82	93	75	209	110	0	0
Tous enrichissements	29.62%	37.35%	30.90	33.04	35.50	93	75	221	110	0	0
Sans coréférence - Focus absent - Rejet des réponses au seuil de 0											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	33.33%	30.52%	32.73	31.87	31.05	76	61	152	128	24	8
Synonymes	31.25%	32.13%	31.42	31.68	31.95	80	64	176	124	27	10
D	32.55%	33.33%	32.70	32.94	33.17	83	67	172	120	28	10
D-Syn	30.97%	33.33%	31.42	32.11	32.83	83	67	185	119	29	11
D-B	32.31%	33.73%	32.58	33.01	33.44	84	68	176	118	29	10
D-EWN-M	33.33%	34.94%	33.64	34.12	34.61	87	71	174	115	28	10
D-Dér	32.18%	33.73%	32.48	32.94	33.41	84	67	177	120	29	11
D-B-EWN-M-Dér	32.72%	35.74%	33.28	34.17	35.09	89	72	183	113	30	11
Tous enrichissements	31.34%	35.74%	32.13	33.40	34.77	89	72	195	113	30	11

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus absent - Rejet des réponses au seuil de 10											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	33.48%	30.52%	32.84	31.93	31.07	76	61	151	129	25	8
Synonymes	31.37%	32.13%	31.52	31.75	31.97	80	64	175	125	28	10
D	32.68%	33.33%	32.81	33.00	33.20	83	67	171	121	29	10
D-Syn	31.09%	33.33%	31.51	32.17	32.86	83	67	184	120	30	11
D-B	32.43%	33.73%	32.68	33.07	33.47	84	68	175	119	30	10
D-EWN-M	33.46%	34.94%	33.75	34.18	34.63	87	71	173	116	29	10
D-Dér	32.31%	33.73%	32.58	33.01	33.44	84	67	176	121	30	11
D-B-EWN-M-Dér	32.84%	35.74%	33.38	34.23	35.12	89	72	182	114	31	11
Tous enrichissements	31.45%	35.74%	32.22	33.46	34.79	89	72	194	114	31	11
Sans coréférence - Focus absent - Rejet des réponses au seuil de 20											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	86.67%	26.10%	59.20	40.12	30.35	65	57	10	138	177	16
Synonymes	84.34%	28.11%	60.24	42.17	32.44	70	60	13	134	200	19
D	81.61%	28.51%	59.46	42.26	32.78	71	62	16	131	196	18
D-Syn	80.00%	28.92%	59.11	42.48	33.15	72	62	18	130	207	20
D-B	78.26%	28.92%	58.35	42.23	33.09	72	63	20	129	197	18
D-EWN-M	80.65%	30.12%	60.39	43.86	34.44	75	66	18	126	196	18
D-Dér	81.61%	28.51%	59.46	42.26	32.78	71	62	16	131	203	19
D-B-EWN-M-Dér	76.77%	30.52%	58.91	43.68	34.70	76	67	23	123	203	19
Tous enrichissements	77.00%	30.92%	59.32	44.13	35.13	77	67	23	124	214	20

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus absent - Rejet des réponses au seuil de 30											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	87.14%	24.50%	57.66	38.24	28.61	61	55	9	140	182	18
Synonymes	86.67%	26.10%	59.20	40.12	30.35	65	57	10	138	208	21
D	82.50%	26.51%	58.00	40.12	30.67	66	59	14	134	203	20
D-Syn	82.72%	26.91%	58.46	40.61	31.10	67	59	14	134	216	22
D-B	78.82%	26.91%	56.88	40.12	30.99	67	60	18	132	204	20
D-EWN-M	81.40%	28.11%	59.02	41.79	32.35	70	63	16	129	203	20
D-Dér	82.50%	26.51%	58.00	40.12	30.67	66	59	14	134	210	21
D-B-EWN-M-Dér	77.17%	28.51%	57.54	41.64	32.63	71	64	21	126	210	21
Tous enrichissements	78.26%	28.92%	58.35	42.23	33.09	72	64	20	127	222	22
Sans coréférence - Focus absent - Rejet des réponses au seuil de 40											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	87.88%	23.29%	56.53	36.83	27.31	58	52	8	143	186	18
Synonymes	87.32%	24.90%	58.16	38.75	29.05	62	54	9	141	212	20
D	82.89%	25.30%	56.96	38.77	29.38	63	56	13	136	207	19
D-Syn	83.12%	25.70%	57.45	39.26	29.82	64	56	13	136	220	21
D-B	79.75%	25.30%	55.75	38.41	29.30	63	56	16	135	210	19
D-EWN-M	82.50%	26.51%	58.00	40.12	30.67	66	59	14	132	209	20
D-Dér	82.89%	25.30%	56.96	38.77	29.38	63	56	13	136	214	20
D-B-EWN-M-Dér	79.52%	26.51%	56.80	39.76	30.58	66	59	17	131	219	21
Tous enrichissements	80.72%	26.91%	57.66	40.36	31.05	67	59	16	132	231	22

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus présent - Sans rejet des réponses sans concordances											
Référence	83.75%	26.91%	58.88	40.73	31.13	67	59	13	136	0	0
Synonymes aveugles	80.90%	28.92%	59.50	42.60	33.18	72	61	17	133	0	0
Syntaxe	84.93%	24.90%	57.30	38.51	29.00	62	55	11	140	0	0
Synonymes	81.48%	26.51%	57.59	40.00	30.64	66	57	15	137	0	0
D	77.27%	27.31%	56.57	40.36	31.37	68	60	20	133	0	0
D-Syn	76.67%	27.71%	56.65	40.71	31.77	69	60	21	132	0	0
D-B	73.40%	27.71%	55.20	40.23	31.65	69	61	25	131	0	0
D-EWN-M	76.60%	28.92%	57.60	41.98	33.03	72	64	22	128	0	0
D-Dér	77.27%	27.31%	56.57	40.36	31.37	68	60	20	133	0	0
D-B-EWN-M-Dér	72.28%	29.32%	55.90	41.71	33.27	73	65	28	125	0	0
Tous enrichissements	73.27%	29.72%	56.66	42.29	33.73	74	65	27	126	0	0
Sans coréférence - Focus présent - Rejet des réponses au seuil de 0											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	86.76%	23.69%	56.62	37.22	27.73	59	53	9	142	5	1
Synonymes	85.14%	25.30%	57.80	39.01	29.44	63	55	11	139	7	2
D	82.28%	26.10%	57.52	39.63	30.23	65	58	14	135	9	3
D-Syn	81.48%	26.51%	57.59	40.00	30.64	66	58	15	134	9	3
D-B	78.57%	26.51%	56.41	39.64	30.56	66	59	18	133	10	3
D-EWN-M	81.18%	27.71%	58.57	41.32	31.91	69	62	16	130	9	3
D-Dér	82.28%	26.10%	57.52	39.63	30.23	65	58	14	135	9	3
D-B-EWN-M-Dér	76.92%	28.11%	57.10	41.18	32.20	70	63	21	127	10	3
Tous enrichissements	78.02%	28.51%	57.91	41.76	32.66	71	63	20	128	10	3

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus présent - Rejet des réponses au seuil de 10											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	86.76%	23.69%	56.62	37.22	27.73	59	53	9	142	5	1
Synonymes	85.14%	25.30%	57.80	39.01	29.44	63	55	11	139	7	2
D	82.28%	26.10%	57.52	39.63	30.23	65	58	14	135	9	3
D-Syn	81.48%	26.51%	57.59	40.00	30.64	66	58	15	134	9	3
D-B	78.57%	26.51%	56.41	39.64	30.56	66	59	18	133	10	3
D-EWN-M	81.18%	27.71%	58.57	41.32	31.91	69	62	16	130	9	3
D-Dér	82.28%	26.10%	57.52	39.63	30.23	65	58	14	135	9	3
D-B-EWN-M-Dér	76.92%	28.11%	57.10	41.18	32.20	70	63	21	127	10	3
Tous enrichissements	78.02%	28.51%	57.91	41.76	32.66	71	63	20	128	10	3
Sans coréférence - Focus présent - Rejet des réponses au seuil de 20											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	86.57%	23.29%	56.09	36.71	27.28	58	52	9	143	6	1
Synonymes	84.93%	24.90%	57.30	38.51	29.00	62	54	11	140	8	2
D	81.82%	25.30%	56.55	38.65	29.36	63	56	14	137	11	3
D-Syn	81.01%	25.70%	56.64	39.02	29.77	64	56	15	136	11	3
D-B	78.05%	25.70%	55.46	38.67	29.68	64	57	18	135	12	3
D-EWN-M	80.72%	26.91%	57.66	40.36	31.05	67	60	16	132	11	3
D-Dér	81.82%	25.30%	56.55	38.65	29.36	63	56	14	137	11	3
D-B-EWN-M-Dér	76.40%	27.31%	56.20	40.24	31.34	68	61	21	129	12	3
Tous enrichissements	77.53%	27.71%	57.02	40.83	31.80	69	61	20	130	12	3

Type	Précision	Rappel	F-m1	F-m2	F-m3	Exact	1 exact	Faux	Sans	Rejet	Rej. exact
Sans coréférence - Focus présent - Rejet des réponses au seuil de 30											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	86.57%	23.29%	56.09	36.71	27.28	58	52	9	143	6	1
Synonymes	86.11%	24.90%	57.73	38.63	29.03	62	54	10	141	9	2
D	81.82%	25.30%	56.55	38.65	29.36	63	56	14	137	11	3
D-Syn	82.05%	25.70%	57.04	39.14	29.80	64	56	14	137	12	3
D-B	78.05%	25.70%	55.46	38.67	29.68	64	57	18	135	12	3
D-EWN-M	80.72%	26.91%	57.66	40.36	31.05	67	60	16	132	11	3
D-Dér	81.82%	25.30%	56.55	38.65	29.36	63	56	14	137	11	3
D-B-EWN-M-Dér	76.40%	27.31%	56.20	40.24	31.34	68	61	21	129	12	3
Tous enrichissements	77.53%	27.71%	57.02	40.83	31.80	69	61	20	130	12	3
Sans coréférence - Focus présent - Rejet des réponses au seuil de 40											
Référence	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	80	0
Synonymes aveugles	0.00%	0.00%	0.00	0.00	0.00	0	0	0	200	89	0
Syntaxe	87.30%	22.09%	54.89	35.26	25.97	55	49	8	146	10	1
Synonymes	86.76%	23.69%	56.62	37.22	27.73	59	51	9	144	13	1
D	82.19%	24.10%	55.45	37.27	28.06	60	53	13	139	15	2
D-Syn	82.43%	24.50%	55.96	37.77	28.50	61	53	13	139	16	2
D-B	78.95%	24.10%	54.25	36.92	27.99	60	53	16	138	18	2
D-EWN-M	81.82%	25.30%	56.55	38.65	29.36	63	56	14	135	17	3
D-Dér	82.19%	24.10%	55.45	37.27	28.06	60	53	13	139	15	2
D-B-EWN-M-Dér	78.75%	25.30%	55.36	38.30	29.28	63	56	17	134	21	3
Tous enrichissements	80.00%	25.70%	56.24	38.91	29.74	64	56	16	135	21	3

Résumé

Cette thèse présente une méthode originale pour identifier et structurer l'information de documents et pour l'interroger. Comme les méthodes linguistiques améliorent les résultats des systèmes actuels, cette approche se base sur des analyses linguistiques et des ressources lexicales. Une analyse grammaticale de haut niveau (morphologique, syntaxique et sémantique) identifie d'abord les éléments d'information et les lie entre eux. Puisque le contexte des requêtes est faible, les textes sont analysés. Puis le contenu des ressources confère aux informations de nombreuses actualisations grâce à des transformations contextuelles : synonymie simple et complexe, dérivations avec adaptation du contexte syntaxique, adjonction de traits sémantiques. . .Enfin, l'interrogation des textes est testée. Une analyse morpho-syntaxique de la question en identifie les éléments d'information et choisit le type de la réponse attendue. Le fragment de texte contenant ces données constitue la réponse à la question.

Mots-clefs : question-réponse, extraction d'information, recherche d'information, désambiguïsation sémantique lexicale, dictionnaire électronique

Abstract

Construct and question the informative structure from a French documentary base

This thesis presents an original methodology to identify and structure information of a French textual base in order to question it. Linguistic techniques in current methods make it possible to improve the results. So we propose a methodology using high-level linguistic analysis (morphology, syntax and word sense disambiguation) to identify each piece of information and to connect them. Because of the lack of context in queries, the texts are analyzed. Then, the information from lexico-semantic resources is used to transform each piece of information in many realizations : simple and complex synonymy, derivation with adaptation to the syntactic context, addition of semantic features and categories. . .We finally tested the questioning method using a morpho-syntactic analysis to identify each piece of information in the question and to determine the semantic type of the required answer ; the passage of the texts containing these data is the answer to the question.

Keywords : question answering (QA), information extraction (IE), information retrieval (IR), word sense disambiguation (WSD), electronic dictionary

Université de la Sorbonne Nouvelle – Paris III ILPGA
Institut de Linguistique et Phonétique Générales et Appliquées
19, rue des Bernardins
75 005 Paris