Formalisation de connaissances à partir de corpus : modélisation linguistique du contexte pour l'extraction automatique de relations sémantiques

Ismaïl El Maarouf

THESE / UNIVERSITÉ DE BRETAGNE SUD
sous le sceau de l'Université européenne de Bretagne
pour obtenir le titre de
DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE SUD
Mention : Sciences et Technologies de l'Information et de la Communication
Ecole doctorale SICMA

présentée par

Ismaïl El Maarouf

Laboratoire de recherche en Informatique
VALORIA

Formalisation de connaissances à partir de corpus

Modélisation linguistique du contexte pour l'extraction automatique de relations sémantiques

Thèse soutenue le 06 Décembre 2011
devant le jury composé de :

Pierre-François Marteau
Professeur Université de Bretagne Sud / *président*

Sophie Rosset
Chargée de recherches CR1 CNRS HDR Limsi / *rapporteur*

Thierry Poibeau
Directeur de recherches CNRS HDR, Lattice / *rapporteur*

Jean-Yves Antoine
Professeur Université François Rabelais de Tours / *examinateur*

Jeanne VILLANEAU
Maître de Conférences Université de Bretagne Sud / Directeur de thèse

RÉSUMÉ DE THÈSE

Les corpus, collections de textes sélectionnés dans un objectif spécifique, occupent une place de plus en plus déterminante en Linguistique comme en Traitement Automatique des Langues (TAL). Considérés à la fois comme source de connaissances sur l'usage authentique des langues, ou sur les entités que désignent des expressions linguistiques, ils sont notamment employés pour évaluer la performance d'applications de TAL. Les critères qui prévalent à leur constitution ont un impact évident, mais encore délicat à caractériser, sur (i) les structures linguistiques majeures qu'ils renferment, (ii) les connaissances qui y sont véhiculées, et, (iii) la capacité de systèmes informatiques à accomplir une tâche donnée.

Ce mémoire étudie des méthodologies d'extraction automatique de relations sémantiques dans des corpus de textes écrits. Un tel sujet invite à examiner en détail le contexte dans lequel une expression linguistique s'applique, à identifier les informations qui déterminent son sens, afin d'espérer relier des unités sémantiques. Généralement, la modélisation du contexte est établie à partir de l'analyse de co-occurrence d'informations linguistiques issues de ressources ou obtenues par des systèmes de TAL. Les intérêts et limites de ces informations sont évalués dans le cadre de la tâche d'extraction de relations sur des corpus de genre différent (article de presse, conte, biographie). Les résultats obtenus permettent d'observer que pour atteindre une représentation sémantique satisfaisante ainsi que pour concevoir des systèmes robustes, ces informations ne suffisent pas.

Deux problèmes sont particulièrement étudiés. D'une part, il semble indispensable d'ajouter des informations qui concernent le genre du texte. Pour caractériser l'impact du genre sur les relations sémantiques, une méthode de classification automatique, reposant sur les restrictions sémantiques qui s'exercent dans le cadre de relations verbo-nominales, est proposée. La méthode est expérimentée sur un corpus de conte et un corpus de presse.

D'autre part, la modélisation du contexte pose des problèmes qui relèvent de la variation discursive de surface. Un texte ne met pas toujours bout à bout des expressions linguistiques en relation et il est parfois nécessaire de recourir à des algorithmes complexes pour détecter des relations à longue portée. Pour répondre à ce problème de façon cohérente, une méthode de segmentation discursive, qui s'appuie sur des indices de structuration de surface apparaissant dans des corpus écrits, est proposée. Elle ouvre le champ à la conception de grammaires qui permettent de raisonner sur des catégories d'ordre macro-syntaxique afin de structurer la représentation discursive d'une phrase. Cette méthode est appliquée en amont d'une analyse syntaxique et l'amélioration des performances est évaluée.

Les solutions proposées à ces deux problèmes nous permettent d'aborder l'extraction d'information sous un angle particulier : le système implémenté est évalué sur une tâche de correction d'Entités Nommées dans le contexte d'application des Systèmes de Question-Réponse. Ce besoin spécifique entraîne l'alignement de la définition d'une catégorie sur le type de réponse attendue par une question.

Mots-clés

Traitement Automatique des Langues, Linguistique, Linguistique de Corpus, Corpus, Sémantique, Relation Sémantique, Extraction d'Information, Entités Nommées, Genre Textuel, Segmentation Discursive, Désambiguïsation, Extraction de Patron, Acquisition de Connaissances à partir de Corpus, Adaptation de Systèmes

Dissertation Summary

Corpora which are text collections selected for specific purposes, are playing an increasing role in Linguistics and Natural Language Processing (NLP). They are conceived as knowledge sources on natural language use, as much as knowledge on the entities designated by linguistic expressions, and they are used in particular to evaluate NLP application performances. The criteria prevailing on their constitution have an obvious, though still delicate to characterize, impact on (i) the major linguistic structures they contain, (ii) the knowledge conveyed, and, (iii) computational systems' success on a give task.

This thesis studies methodologies of automatic extraction of semantic relations on written text corpora. Such a topic calls for a detailed examination of the context in which a given expression holds, as well as for the discovery of the features which determine its meaning, in order to be able to link semantic units. Generally, contextual models are built from the co-occurrence analysis of linguistic informations, drawn from resources and NLP tools. The benefits and limits of these informations are evaluated in a task of relation extraction from corpora belonging to different genres (press article, fairy tale, biography). The results show that these informations are insufficient to reach a satisfying semantic representation as well as to design robust systems.

Two problems are particularly addressed. On the one hand, it seems indispensable to add informations related to text genre. So as to characterize the impact of genre on semantic relations, an automatic classification method, which relies on the semantic restrictions holding between verbs and nouns, is proposed. The method is experimented on a fairy tale corpus and on a press corpus.

On the other hand, contextual models need to deal with problems which come under discourse surface variation. In a text, related linguistic expressions are not always close to one another and it is sometimes necessary to design complex algorithms in order to detect long dependencies. To answer this problem in a coherent manner, a method of discourse segmentation based on surface structure triggers in written corpora, is proposed. It paves the way for grammars operating on macro-syntactic categories in order to structure the discursive representation of a sentence. This method is applied prior to a syntactic analysis and its improvement is evaluated.

The solutions proposed to these problems help us to approach Information Extraction from a particular angle : the implemented system is evaluated on a task of Named Entity correction in the context of a Question-Answering System. This specific need entails the alignment of a category definition on the type of answer expected by the question.

Key Words

Natural Language Processing, Linguistics, Corpus Linguistics, Corpus, Semantics, Semantic Relations, Information extraction, Named Entities, Text Genre, Discourse Segmentation, Disambiguation, Pattern Extraction, Corpus-driven Knowledge Acquisition, System Adaptation

REMERCIEMENTS

Qu'il me soit ici permis de remercier les personnes qui ont contribué directement ou indirectement aux recherches que j'ai menées durant ces trois ans à l'université de Bretagne-Sud.

Tout d'abord j'aimerais remercier Thierry Poibeau et Sophie Rosset d'avoir accepté de rapporter mon travail. Leurs critiques m'ont fourni matière à approfondir et revoir des points importants auxquels touche cette thèse qui constitueront des pistes de recherches fructueuses à venir. L'intérêt qu'ils ont montré pour ce travail m'encourage à penser que des discussions collaboratives pourront voir le jour. J'ai particulièrement une dette considérable envers Sophie, qui m'a abondamment procuré données et systèmes, mais également encouragé et orienté dans mes recherches. Cette thèse lui doit énormément.

Je remercie également Pierre-François Marteau et Jean-Yves Antoine d'avoir témoigné leur intérêt envers mes recherches en ayant accepté d'examiner ce travail. Jean-Yves Antoine qui était aussi membre de mon comité de thèse, a fait preuve d'une probité exemplaire à mon égard.

Ensuite, il me revient de remercier la personne qui a scrupuleusement et méthodiquement encadré mon travail, ma directrice de thèse, Jeanne Villaneau. Les ressources scientifiques et humaines qu'elle a déployées, sa confiance et son enthousiasme ont considérablement enrichi mon travail et aiguisé mon esprit scientifique. J'espère également que des occasions nous seront données pour poursuivre nos collaborations.

Je remercie toute l'équipe du laboratoire Valoria, au sein duquel j'ai réalisé ce travail. J'ai aussi une pensée pour les membres du laboratoire HCTI qui m'a accueilli aux débuts de ma thèse et particulièrement Rémi Le Marc'hadour, pour avoir suivi et encouragé ma progression.

Il n'est pas possible de citer toutes les personnes de l'université Bretagne-Sud et de Rennes 2, des laboratoires de recherches et des services associés qui ont contribué à l'atmosphère générale. Merci à Amel, André, Christophe, Chrystel, Delphine, Dominique, Farida, Gael, Gildas, Hiroyasu, Marc, Marie, Nolwenn, Perig, Radia, Sébastien, Sylvain, Xavier, et à tant d'autres pour leur bonne humeur et leur amitié.

Au-delà de cette université, de nombreuses personnes m'ont conseillé dans mes recherches, un merci particulier aux innombrables toulousains, le CLLE_ERSS avec Cécile, Clémentine, Didier, Fanny, Ludovic et Marie-paule, l'IRIT avec Patrick Saint-Dizier, Orange Labs, avec Émilie, Johannes, Malek, Olivier, le CENTAL, avec Jean-Léon, Patrick, Richard et Thomas et j'en manque.

La distance n'aura pas eu raison de l'attention que m'ont porté mes amis et ma famille, qui ont parfois bravé les tempêtes bretonnes pour venir me voir dans mon trou de campagne. Un merci particulier à Anna, Belkacem, Brigitte, Bruno, Fabien, Florian, Jean-Luc, Marie-Paule, Nathalie, Sami, Tarik et Thomas.

J'ai une pensée toute particulière pour ma mère Danielle, mon père Abdeslem, ma belle-mère Souad, mes frères et soeurs, en commençant par le plus vieux, Khalil, Mounir, Samir, Hanane, Safaa, Imad, Youssef et Hamza, mes tantes, oncles, cousins, français et marocains.

Un merci également à ma belle-famille, Christian, Dominique, Gwenaëlle, Marion, Virginie, et leurs adorables bouts de chou et compagnons respectifs.

Je réserve ces dernières lignes à la femme de ma vie, Nathalie, dont les qualités scientifiques et humaines sont innombrables, d'une abondante générosité, pour son amour.

'Tis but thy name that is my enemy;
Thou art thyself, though not a Montague.
What's Montague? it is nor hand, nor foot,
Nor arm, nor face, nor any other part
Belonging to a man. O, be some other name!
What's in a name? that which we call a rose
By any other name would smell as sweet;
So Romeo would, were he not Romeo call'd,
Retain that dear perfection which he owes
Without that title. Romeo, doff thy name,
And for that name which is no part of thee
Take all myself.

Shakespeare, Romeo and Juliet, Act II Scene 2

Table des matières

Sémantique et texte : l'épreuve des contes

Chapitre 3. Analyse collocationnelle

Chapitre 4. Patron sémantique en corpus

Sémantique pour l'extraction d'information

Chapitre 5. Extraction d'information et Entités Nommées

Chapitre 6. Acquisition de cadre d'EI

Segmentation discursive de surface

Chapitre 7. Segmentation de surface

Liste des illustrations

INDEX DES TABLES

Liste des abréviations utilisées

ACE	Automatic Content Extraction (Campagne)
BNC	British National Corpus
BSO	Brandeis Semantic Ontology
CE	Chunk Entité (Type de Modèle)
CEM	Chunk Entité Mixte (Type de Modèle)
CF	Chunk Forme (Type de Modèle)
COBUILD	Collins Birmingham University International Language Database
CPA	Corpus Pattern Analysis (construction de patron sémantique)
DDR	Descripteur de Recherche
EAT	Expected Answer Type (Type de Réponse Attendue)
EI	Extraction d'Information
EN	Entité Nommée
ENAMEX	Entity Name Expression (Expression d'entité nommée)
EnCor	Système de CORrection d'Entités Nommées
Entité-R	Catégorie fournie par le système Ritel-nca
Ester2	Campagne Ester2
GPE	Géopolitical Entity (Entité Géopolitique)
IM	Information Mutuelle (Mesure)
LCI	Linguistique de Corpus Informatisé
MAX	Meilleur score des Patrons (Type de Score de Segment)
MEAN	Moyenne des scores des Patrons (Type de Score de Segment)
MUC	Message Understanding Conference (Campagne)
Npr	Nom Propre
NUMEX	Numérical Expression (Expression de Valeur)
PROBA	Probabilité (Mesure)
PROD	Produit des scores des Patrons (Type de Score de Segment)
R ou Rnc	Système RITEL-nca
RCEN	Reconnaissance et Classification d'Entité Nommée
RI	Recherche d'Information
Semeval	Semantic Evaluation (Campagne)
SQR	Système de Question-Réponse
TAL	Traitement Automatique des Langues
Tlfi	Trésor de la Langue Française Informatisé
TREC	Text Retrieval Conference
UES	Unité Étendue de Sens

INTRODUCTION

La révolution numérique est derrière nous et ses effets sur les pratiques culturelles se manifestent dans tous les domaines, politique, juridique, médical, éducatif, commercial et bien sûr, scientifique. En 2008, plus de 65 français sur 100 possédaient déjà un ordinateur et 56 avaient accès au plus important réseau d'échange d'information mondial, Internet[1]. Une quantité d'information est produite chaque jour, diffusée, traitée, consultée, et conservée sur support informatique. Au milieu de ce vaste océan numérique, la conception de technologies d'accès à l'information est devenue une priorité. Parmi les vecteurs d'information, le texte occupe toujours une place centrale. Le domaine de la Recherche d'Information a connu un important développement et atteint une maturité qui permet à chaque utilisateur de brasser des milliers de milliards de textes et d'y retrouver des documents pertinents. Les moteurs de recherche mesurent la pertinence d'un document en fonction de la fréquence d'apparition des mots-clés d'une requête dans un texte, ou d'autres paramètres comme le nombre de sites pointant vers ce document dans le cas du web. Néanmoins, chaque utilisateur le sait, plus la recherche d'information se complexifie, plus les limites des moteurs de recherche apparaissent : par exemple, tel moteur ne traite pas les accents, il ne « connaît » pas la différence entre le singulier et pluriel, il ne « comprend » pas que tel mot est employé dans tel sens et enfin il aurait pu « savoir » que ce mot pouvait être reformulé de telle manière. Les moteurs de recherche sont d'une utilité précieuse, mais on voudrait pouvoir aller plus loin, en d'autres termes que la recherche soit plus structurée et plus sémantique, qu'elle intègre des connaissances linguistiques. Le domaine du Traitement Automatique des Langues (TAL) occupe une place privilégiée pour explorer ces frontières où la linguistique et l'informatique se rencontrent.

Le TAL est orienté vers le développement d'applications dont la diversité des objectifs se traduit en autant de sous-domaines : la traduction automatique, le résumé automatique de document, la correction automatique et la recherche et l'extraction d'information. Les performances des systèmes sont évaluées par rapport à une tâche sur des données langagières. Les approches sont très variables mais relèvent toutes d'une ingénierie de la langue reposant sur la définition de méthodes quantitatives et l'exploitation de ressources, comme les lexiques. Le domaine semble clairement circonscrit et les applications se sont progressivement diversifiées.

La diversité et la complexité des tâches auxquelles sont soumis les systèmes supposent d'identifier les composants qui peuvent être réutilisés. Comme le remarque J.-Y. Antoine, cet objectif de généricité du système soulève des questions fondamentales qui participent de l'auto-critique du TAL :

> « D'un point de vue ingénierique, la portabilité et la réutilisabilité des systèmes constituent des facteurs de qualité essentiels. Leur importance économique n'est d'ailleurs pas à démontrer. La généricité ne saurait être cependant limitée à ces aspects purement technologiques. Au contraire, elle interroge l'ingénierie des langues dans ses fondements et ses pratiques actuelles. » [Antoine, 2003 : 50]

1 [Donnat, 2009 : 28]

Cette réflexion nous invite à poser tout particulièrement les questions suivantes : les systèmes peuvent-ils être exploités dans d'autres sous-tâches de TAL que celles pour lesquelles ils ont été conçus ? Que partagent, par ailleurs ces sous-tâches ? Quels problèmes ces systèmes permettent-ils de résoudre et quels phénomènes langagiers sont-ils capables de modéliser ? Enfin, sur quels textes peut-on les utiliser ?

Toutes ces questions renvoient à la complexité de l'objet-même du TAL : le langage, à travers ses incarnations dans les langues et les textes. En effet, si les applications du TAL peuvent faire l'économie d'une modélisation linguistique à partir du moment où la tâche effectuée demeure relativement simple, dès que l'objectif visé se complexifie et nécessite l'articulation de différentes tâches complexes, le TAL montre ses limites.

Pour surmonter ces difficultés, il paraît légitime de se tourner vers les disciplines scientifiques qui s'intéressent au même objet d'étude, et plus particulièrement vers la linguistique. Les recherches en linguistique s'attachent à décrire la diversité des phénomènes langagiers et ont abouti à la formalisation des systèmes théoriques pour modéliser cette complexité.

Nous nous interrogeons, à la suite de JY Antoine, sur les bénéfices que pourrait tirer le TAL de l'examen de ces recherches en linguistique.

Cette thèse a pour objectif d'explorer ces passerelles dans le cadre de l'extraction automatique de relations sémantiques. Analyser les relations sémantiques, c'est mettre un pied sur le terrain de la pensée et l'autre sur celui du langage. En effet, l'extraction sémantique automatique s'inscrit dans le domaine de la compréhension automatique des langues (« Natural Language Understanding » ; [Sabah, 2010]) et relève plus spécifiquement du domaine de l'extraction d'information. Celle-ci consiste à extraire des catégories et des relations sémantiques ciblées répondant à un objectif d'information spécifique. Elle vise une compréhension partielle pour permettre un accès au contenu localisé dans les documents.

Or, A. Nazarenko souligne le paradoxe de l'acquisition de connaissances qui s'effectue à partir de textes :

> « comprendre mobilise des connaissances, mais il faut comprendre pour acquérir des connaissances » [Nazarenko, 2004 : 9]

Ainsi, mieux décrire, saisir, assimiler la connaissance que recèle le texte, permettra de mieux mobiliser ces informations pour acquérir de nouvelles connaissances. Mais, face à un objet aussi complexe que le langage, quelles connaissances peuvent alimenter ce « cercle vertueux » [*ibid.*] et comment y contribuent-elles ?

Que doit-on savoir sur le langage, que doit-on connaître sur les textes, quels critères sont déterminants pour extraire des relations sémantiques ?

Une relation sémantique ne s'établit pas entre deux mots mais entre les concepts auxquels ils sont associés. L'identification d'une relation sémantique suppose des connaissances sur les sens des mots, sur ce « contenu » que le langage véhicule, mais aussi, et surtout, des connaissances sur l'usage des mots, sur les structures linguistiques qui portent ce contenu.

Notre problématique sera justement d'identifier le type de connaissances mises à profit dans l'acquisition de relations sémantiques. Afin de caractériser les difficultés auxquelles fait face cette entreprise, nous ne nous contenterons pas de lister ces connaissances, elles seront évaluées de manière à déterminer la part que chacune d'elles prend.

Trois directions de recherche distinctes nous permettront de contribuer à la résolution de cette problématique :

L'acquisition de connaissances à partir de corpus.

Un corpus est une collection de textes réunis dans un objectif spécifique. Il constitue la ressource majeure de connaissances pour les systèmes de TAL. Nous nous interrogerons sur le type d'informations que l'on peut en extraire à partir d'une analyse du contexte lexical et textuel.

L'applicabilité des modèles linguistiques au texte.

L'approche proposée dans cette thèse est linguistiquement motivée. Plusieurs modèles linguistiques seront étudiés. Nous chercherons à savoir quel éclairage ils portent sur les problèmes d'extraction et de quelle manière il sera possible d'intégrer ces informations complémentaires. Cela invitera à dépasser d'éventuelles limites de ces modèles.

L'adaptabilité des systèmes d'extraction sémantique.

Les systèmes d'extraction mobilisent des sources de connaissances diverses, comme les corpus, les lexiques et les grammaires. Nous nous intéresserons à l'exploitation et à l'articulation de ces ressources dans un contexte donné afin d'analyser la manière dont chacune d'elles permet l'extraction dans ce contexte précis. Cela nous permettra d'envisager des modalités de transposition de ces ressources pour obtenir des résultats tout aussi corrects mais dans un contexte d'application différent.

La contribution principale de cette thèse repose sur la définition d'un modèle de segmentation discursive de surface pour l'extraction de relation sémantique, modèle qui vise à répondre à deux problèmes majeurs en TAL de manière générale et en Extraction d'Information plus spécifiquement :

La détection de relation à longue distance.

Un des problèmes majeurs auxquels font face les systèmes de TAL est la capacité à détecter des relations entre des unités qui sont séparées par des informations multiples et variées. La segmentation discursive permet de raisonner sur des unités plus larges que le mot et de mettre en relation des éléments distants. Le système sera évalué sur la détection de la relation Sujet.

La désambiguïsation d'Entités Nommées.

Les Entités Nommées constituent les noyaux d'information primaires en extraction d'information. Leur classification correspond à une tâche de catégorisation sémantique qui nécessite de prendre en compte leur contexte d'emploi. Pour les trois catégories d'Entités Nommées principales, nous proposerons d'évaluer la qualité de la correction effectuée par notre système afin d'améliorer les résultats de systèmes préexistants conçus pour une application donnée, en l'élargissant à des contextes d'application différents.

L'exposé de ce travail de recherche en cinq volets permettra de mettre en évidence une contribution possible de la linguistique à l'application du TAL sur des tâches complexes :

1. Le premier volet a l'objectif de définir les concepts-clés qui seront employés dans cette thèse. L'approche linguistique y est détaillée sur le plan théorique et pratique.
    - Le premier chapitre introduira la sémantique. Nous y définirons les phénomènes majeurs à travers deux paradigmes d'analyse, la sémantique référentielle et la sémantique textuelle.
    - Nous enrichirons ce modèle en tenant compte des travaux menés en linguistique de corpus. La quantification des données linguistiques sera plus spécifiquement examinée au travers de l'approche de J. McH. Sinclair. Nous analyserons notamment les concepts d'usage et de collocation.
2. L'apport de ces modèles linguistiques sera évalué dans le second volet consacré à l'analyse sémantique d'un corpus de contes. Nous nous intéresserons aux problèmes de modélisation sémantique du contexte en étudiant plusieurs types d'information linguistique.
    - Le chapitre 3 s'intéressera aux relations sémantiques obtenues à partir d'une analyse collocationnelle et à l'apport de la syntaxe dans la sélection des collocations.
    - Le chapitre 4 poursuivra cette étude en présentant des méthodes d'extraction sémantique basées sur les ontologies. Nous chercherons à identifier plus spécifiquement les dépendances entre catégories sémantiques et genre de texte.
3. Le troisième volet présente le domaine de l'Extraction d'Information avec pour objectif d'identifier et caractériser les principales difficultés rencontrées.
    - Le chapitre 5 présentera les Entités Nommées ainsi que les systèmes d'extraction.
    - Le chapitre 6 s'intéressera à l'extraction d'information biographique par des systèmes symboliques.
4. Le quatrième volet présentera notre méthode de segmentation discursive pour répondre aux limites identifiées dans le troisième volet.
    - Le chapitre 7 définira le modèle de segmentation, décrira le segmenteur, et caractérisera les données qu'il permet d'obtenir : les segments.
    - Le chapitre 8 présentera la conception d'une grammaire de segments pour l'identification et l'association de structures discursives.
    - L'apport de cette grammaire sera évalué chapitre 9 dans le cadre d'une tâche de détection de relations syntaxiques à longue distance.
5. Le dernier volet de cette thèse exploitera le modèle de segmentation discursive pour la conception d'un correcteur d'Entités Nommées.
    - Le correcteur sera présenté chapitre 10. Il s'appuie sur l'extraction de patrons sémantiques à partir de corpus pour classifier les catégories d'Entités Nommées.
    - La qualité de cette correction sera évaluée chapitre 11.

PREMIER VOLET

SÉMANTIQUE THÉORIQUE ET PRATIQUE : CONCEPTS, MÉTHODES ET ENJEUX

1. Sémantique

L'étude scientifique du langage et des langues est le domaine privilégié de la linguistique. Elle est tout à la fois affaire de théorie et de pratique. La linguistique n'est pas monolithique, au contraire, elle embrasse une variété de mouvements, d'écoles, de théories et de modèles. Elle partage son terrain avec la psychologie, la philosophie, ou encore la sociologie. Le langage est un objet complexe, comme le reflète la variété de points de vue linguistiques ou « plans d'analyse » : la phonologie, par exemple, concerne l'étude des sons et de leurs représentations, la morphologie étudie de la formation des mots (unités lexicales et morphèmes), etc. Le plan d'analyse principal qui nous concernera dans cette thèse est la *sémantique*, l'étude du sens et de la signification.

Nous proposons d'introduire le sujet en décrivant deux perspectives théoriques qui nous permettront d'engager une réflexion scientifique sur l'étude sémantique du langage et plus particulièrement sur les rapports entre mot et texte. En premier lieu, nous aborderons le champ de la sémantique lexicale, ou sémantique du mot, en nous appuyant sur le paradigme référentiel (1.2). Ensuite, nous examinerons les concepts-clés de la sémantique textuelle, pour tenter de discerner le genre de phénomènes sémantiques que génère un texte (1.3). Ces jalons posés nous amèneront à proposer des points de rencontre pour une sémantique linguistique qui cherche à rendre compte de ces deux courants, souvent opposés (1.4). La première partie de ce chapitre établit les principes préalables à toute analyse sémantique linguistique.

1.1. La question de l'intégration de la sémantique dans la linguistique

1.1.1. Autonomie de la sémantique

La légitimité de la sémantique en linguistique ne va pas de soi. Certains linguistes se refusent à donner une existence à quelque chose d'aussi insaisissable que le sens (voir 1.1.2.), alors que d'autres considèrent que le sens est au cœur du langage et qu'il transcende ainsi tout plan d'analyse. Nous commencerons par ces derniers. A. Wierzbicka, par exemple, formule la thèse suivante :

i. le langage est un outil pour exprimer du sens [Wierzbicka, 1992 : 3].

Dans ce cas, une des tâches de la linguistique est de :

ii. décrire comment les langues créent et véhiculent du sens

Pour décrire ce sens et, à moins de s'intéresser à des problématiques d'acquisition du langage, il faut présupposer l'existence d'un lexique, *i.e.* d'un système de connaissance structuré permettant la reconnaissance et la manipulation d'unités linguistiques. La fonction primaire d'un lexique est, à la manière d'un dictionnaire, de livrer le sens de chaque unité. Il doit avoir un substrat mental (chaque individu est doté d'une compétence lexicale) et une dimension sociale (les individus d'une communauté donnée partagent un lexique commun). L'étude du lexique sur le plan sémantique porte le nom de sémantique lexicale (par opposition à la morphologie lexicale par exemple).
D'une part, il s'agit d'étudier les différents sens d'un mot, de les structurer et de proposer des moyens pour les caractériser, voire les formaliser (figure 1.1).

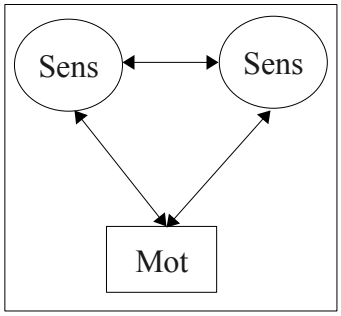


FIG 1.1 – Rapports entre sens d'un mot

D'autre part, il est également nécessaire d'étudier les rapports de sens qu'entretiennent les mots, les différents degrés de synonymie, d'antonymie, etc. (figure 1.2).

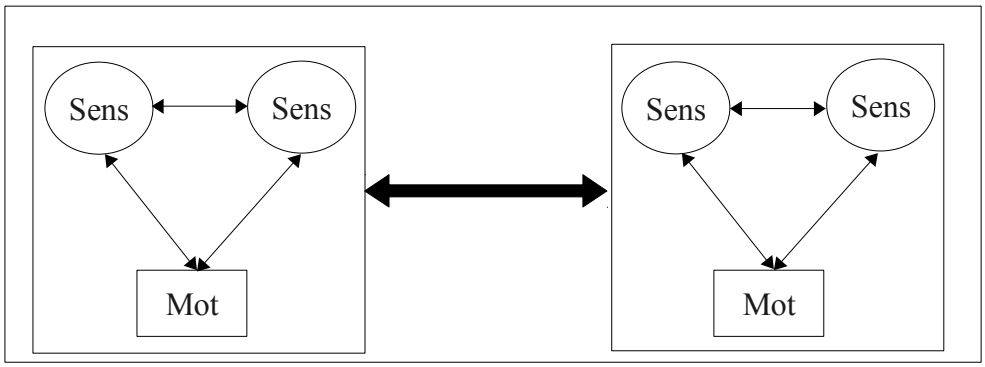


FIG 1.2 – Rapports de sens entre mots

Étant donnée la place qu'occupe le sens dans un tel domaine, il est scientifiquement nécessaire de proposer une définition de ce qui est mis en relation, autrement dit le sens. Cette question peut être abordée, comme nous le verrons, de différents points de vue : selon la célèbre formule du linguiste F. de Saussure, « le point de vue crée l'objet » [de Saussure, 1916 : 23].

Le système lexical s'oppose en principe au système grammatical. L'étude de la grammaire, la syntaxe, caractérise les mots en fonction de leurs propriétés combinatoires, isole des règles de composition, et identifie les structures majeures d'une langue. L'étude de la grammaire peut être menée indépendamment de critères sémantiques, dans le sens où les catégories et structures sont établies indépendamment de la nature des mots (oui, mais quelle est-elle ?). En effet, tous les mots ne fonctionnent pas de la même manière, mais certains mots, si : les catégories de déterminant ou de verbe ont, par exemple, des propriétés syntaxiques bien distinctes.

Nous devons préciser quelques bémols à cette stricte répartition des tâches :

- Si la syntaxe limite généralement son étude au domaine de la phrase, rien n'empêche de considérer que ses attributions s'appliquent à tout type de phénomène de composition linguistique linéaire, du niveau morphologique jusqu'au niveau discursif.
- L'existence de modèles de grammaire sémantique comme la Grammaire Cognitive [Langacker, 1987] pousse à croire que même les catégories syntaxiques de déterminant ou de nom sont sémantiquement motivées.
- Certaines théories postulent une continuité entre les pôles lexical et grammatical [*ibid.*] [Givón, 1985].
- Enfin certaines entreprises de formalisation du lexique incluent des propriétés grammaticales. J. Pustejovsky [Pustejovsky, 1998] propose des paradigmes lexico-conceptuels, agrégats richement structurés de contraintes sémantiques permettant de prédire/expliquer leur combinatoire grammaticale.

Pour ces modèles qui choisissent d'inclure explicitement (au moins) un peu de sémantique dans leur lexique ou un soupçon dans leur syntaxe, se pose la question de l'autonomie de la sémantique vis-à-vis de la linguistique : ces catégories du lexique ou de la grammaire sont-elles propres aux mots (« attachées », « contenues », pour prendre des métaphores concrètes) ou relèvent-elles d'un niveau distinct vers lequel les mots « pointent » ? Autoriser l'existence d'un niveau de représentation sémantique distinct, hors de la linguistique (pas nécessairement universel), c'est lui attribuer une autonomie. Après tout, les mots ne sont pas les seuls « outils » permettant de véhiculer du sens : on peut transmettre une même idée en utilisant des gestes ou des mots. Ceci porte à croire

que la sémantique est autonome, voire, préside la linguistique. Le sens serait alors un objet structuré que l'on pourrait étudier dans sa cohérence interne (ses propres unités et structures). On comprend par là même également qu'il puisse exister des théories linguistiques qui ne prétendent pas étudier le sens. Mais le sens n'est-il pourtant pas toujours évoqué, manipulé, structuré par des signes ? En ce cas, une sémantique linguistique devra tenir compte des deux niveaux : linguistique et sémantique, forme et sens. Cette interdépendance permet de définir une sémantique linguistique, intermédiaire, qui recherche des correspondances entre les structures linguistiques et les structures sémantiques (figure 1.3).

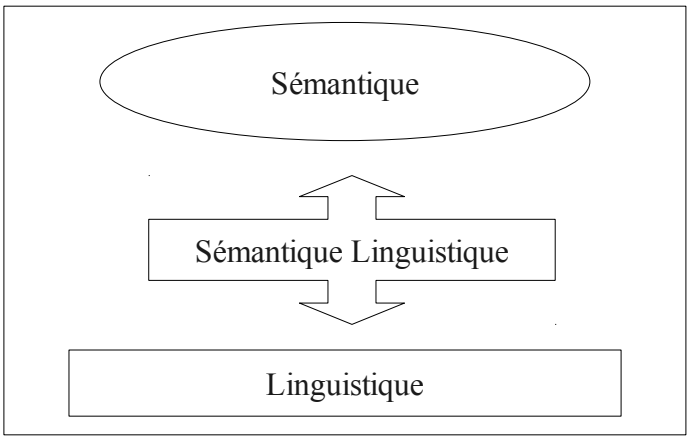


FIG 1.3 – Place de la sémantique linguistique

A priori, le sens auquel on s'intéresse serait indissociable de son matériau support, la langue. Chaque langue « possèderait » une représentation spécifique du sens, même si l'existence de modèles sémantiques universels n'est pas exclue.

Au fond, qu'entend-t-on par *sens* ? Pour tenter d'y répondre, admettons tout d'abord que le sens se caractérise par le point de vue adopté et qu'il peut donc recouvrir différents aspects : culturel, sociologique, réel, conceptuel, perceptuel, etc. Il n'y a pas, de notre point de vue, de sémantique sans référentiel interprétatif. Ce référentiel est d'autant plus important qu'il constitue un argument de stabilité pour le linguiste face à la critique de l'introspection (voir le problème de la preuve évoqué au chapitre 2) : l'interprétation, en tant qu'elle s'ancre dans un référentiel, n'est plus uniquement propre à celui qui la profère, mais est d'une certaine manière contrôlée, contrainte par le référentiel, quel qu'il soit. Par exemple, dire qu'un énoncé (complet) comme *le chien vole* n'est pas sémantiquement valide parce que dans la réalité, les entités désignées par le mot *chien* n'ont pas la qualité de volatile, consiste à utiliser la réalité concrète comme référentiel pour décrire le sens des mots et non à supputer une interprétation personnelle. En d'autres termes, cela revient à définir le sens des mots à partir des relations qui existent dans la réalité, et à considérer valides les énoncés qui traduisent correctement (peuvent être observés dans) la réalité.

Si l'on peut concevoir de nombreuses dimensions au sens, les recherches actuelles en sémantique linguistique nous proposent deux perspectives fondamentales : lier la linguistique au texte et lier la linguistique au réel. C'est ainsi que deux linguistes français contemporains, F. Rastier et G. Kleiber, les dessinent.

Rastier [Rastier, 2001] défend pour sa part la « tradition rhétorico-herméneutique », qui étend l'étude du sens aux rapports qu'entretiennent les mots dans le contexte d'un texte et à ceux qu'entretiennent les textes qui sont ancrés dans des pratiques socio-culturelles. Il formule un

avertissement aux linguistes : le langage peut être influencé par ses milieux de production et d'interprétation. Un texte s'inscrit toujours dans une histoire, un contexte, une pratique culturelle et c'est relativement à ces catégories « humaines » qu'il fonctionne. Nous aurons l'occasion de développer cette perspective en 1.3.

Kleiber [Kleiber, 1999a] défend le paradigme référentiel en faisant l'hypothèse que le langage est principalement tourné vers le réel. Le langage permet de nommer, désigner, catégoriser les objets du monde. Cette conception est solidement ancrée dans l'histoire de la linguistique, et, bien qu'elle ait essuyé des critiques, est toujours la plus répandue (cf. *supra* 1.2).
Les chercheurs qui se réclament de ces perspectives s'opposent, mais ces dernières nous paraissent complémentaires et toutes deux essentielles dans une perspective multi-dimensionnelle du sens linguistique, ce que nous aurons l'occasion de discuter en 1.4.

1.1.2. Subjectivité et sens

Aussi paradoxal que cela puisse paraître, la problématique du sens a longtemps été écartée de l'analyse linguistique, se traduisant par le rejet de la thèse (i) (p7). Cette position est souvent associée à une volonté d'objectivité nécessaire pour établir la linguistique comme discipline scientifique, le sens introduisant une dimension subjective. L. Bloomfield [Bloomfield, 1933], par exemple, en considérant le sens comme relevant du mental, du psychologique, en arrive à la conclusion suivante :

> « The study of language can be conducted without special assumptions so long as we pay no attention to the meaning of what is spoken. » [*ibid.* : 75]

Il s'agit d'une position de type structuraliste : la langue est un objet structuré, un système dont on doit identifier les régularités formelles, voire sous-jacentes. Introduire le concept de sens, c'est malmener l'objectivité scientifique. Ce courant a connu un essor particulier en linguistique, on peut lui rattacher certaines conceptions de Z. Harris, ou de N. Chomsky. Là où certains privilégieront une analyse appliquée, empirique (Harris et Bloomfield), d'autres chercheront à caractériser le système abstrait sous-jacent ou commun qui permet l'exercice du langage (la compétence pour Chomsky). Tous partagent néanmoins une même suspicion vis-à-vis de l'étude du sens, mettant ainsi à mal l'établissement d'une sémantique.

Le problème que rencontrent toutes les théories du sens est ainsi celui de la subjectivité de l'interprétation : le sens est intangible et instable donc inconnaissable et indescriptible. Pour critiquer cette position, il est possible de livrer un premier contre-argument, de nature pragmatique, consistant à faire observer que, sans sens, la communication n'a plus lieu d'être :

> « La chose remarquable est cependant que ceux qui remettaient radicalement en question la stabilité intersubjective du sens ne soulevaient jamais la question de savoir comment leurs propres affirmations pouvaient avoir un sens stable et connaissable. Si le sens était radicalement indéterminé, comment pouvait-il en être autrement pour le méta-langage de la linguistique ? » [Larsson, 2008 : 29]

Tout modèle linguistique, pour être compréhensible, doit présupposer l'existence du sens et son intersubjectivité, c'est la conclusion qui semble s'imposer. Toute catégorie nécessite d'être interprétée et partagée. On peut ainsi formuler les thèses suivantes :

iii. le sens existe

iv. le sens doit être partagé

De plus, accepter que le sens existe signifie supposer l'existence d'une forme de rationalité dans l'objet d'étude. *A contrario*, chercher à expliquer un phénomène revient à supposer l'existence du sens. Que cette rationalité réside dans le modèle que l'on cherche à appliquer à l'objet ou dans l'objet lui-même (la distinction nous paraît importante), le résultat est le même : l'objet est *considéré* comme rationnel, c'est-à-dire compréhensible. En linguistique, cela implique par exemple qu'on suppose qu'un énoncé véhicule du sens, et qu'en tant qu'il en véhicule un, ce sens peut être décrit et formalisé. D'ailleurs, ce n'est *que* parce que l'on suppose l'existence du sens, que l'on peut juger, identifier des énoncés insensés (qui existent).

Ces considérations ne résolvent qu'une partie du problème puisqu'elles n'expliquent pas encore en quoi consistent les unités de sens, ce qu'elles désignent. Une des positions communes est de considérer que les mots sont des étiquettes des choses que l'on rencontre dans la réalité : décrire le sens des mots revient à établir une liste de termes d'une taxonomie. Cette position est fortement critiquée comme une conception nomenclaturiste de la langue depuis au moins de Saussurre [de Saussure, 1916 : 97] : elle court-circuite la sémantique linguistique et réduit la sémantique à la réalité. Ce type de conception ne revient pas à utiliser un référentiel interprétatif pour décrire le sens mais à *remplacer* le sens par ce référentiel. La sémantique référentielle est souvent créditée d'une telle conception, mais nous allons voir qu'il est tout à fait possible de conserver la réalité comme référentiel d'interprétation dans le cadre d'une sémantique linguistique.

1.2. Le paradigme référentiel

Nous avons précédemment observé qu'il n'y a pas de contradiction de principe entre la recherche d'une motivation extra-linguistique (hors de la linguistique) au mot et l'existence d'une sémantique linguistique. La sémantique référentielle prend pied dans un vaste corps de travaux sur les liens entre langage et réalité, que Kleiber résume sous la dénomination « paradigme référentiel ». Pour introduire le sujet, nous présenterons deux modèles opposés de la référence.

1.2.1. La référence : Logique ou Psychologique ?

Le paradigme référentiel prend pour point de départ l'idée que l'une des fonctions inaliénables du langage est de désigner les entités et les événements qui prennent place dans le monde dans lequel nous vivons, communément dénommé « réalité » : cette fonction est la *référence*. En laissant momentanément de côté le problème de la réalité de cette réalité, on peut tout d'abord constater que le langage nous permet de nous situer dans la réalité, vis-à-vis d'objets, d'entités, de situations, spécifiques ou génériques. Ce constat de départ ne peut se suffire d'un modèle duel du signe linguistique tel que l'a proposé de Saussure (forme-sens ou signifiant-signifié) parce qu'il doit expliquer comment un signe peut représenter un objet.

Les premiers travaux modernes sur le lien entre mot et objet abordent le problème de la référence en termes de conditions de vérité. G. Frege, considéré comme le père de la Logique Formelle, définit le sens en termes des conditions de vérité que doit satisfaire (vérifier) un énoncé pour être vrai, *i.e.* pour correctement désigner une situation donnée. Pour y parvenir, il propose que la signification d'une expression tienne en un double aspect, son sens et sa dénotation. La dénotation est l'objet unique désigné par une expression : elle est vraie si et seulement si une expression désigne un tel individu. Le sens est le « mode de donation » de la dénotation d'une expression, ou la forme par laquelle on désigne un référent. L'exemple qu'il donne pour illustrer cette distinction est *l'étoile du soir* et *l'étoile du matin* qui dénotent tous deux le même référent, mais qui le présentent différemment. Ce sont deux « modes de donation » différents d'une même dénotation. Ces deux expressions ne sont donc pas strictement synonymes même si elles ont la même dénotation (elles désignent le même objet). Mais tous les « sens » ne peuvent prétendre au statut de mode de donation et la définition qu'il fait du sens (par exclusion et différences) est quelque peu caduque. Frege ajoute que le sens doit être objectif et il le distingue par là des représentations subjectives que chaque individu associe à un mot. En fin de compte, le sens doit correspondre aux propriétés que l'on assigne à un objet en utilisant une expression. Comprendre, c'est avoir la compétence d'identifier la dénotation d'une expression. Cette analyse fait du sens un intermédiaire « secondaire » pour ce qui est véritablement visé : l'objet ou la dénotation. Mais les langues se prêtent mal à de telles rigidités et Frege en est tout à fait conscient :

> « Le lien régulier entre le signe, son sens et sa référence, est tel qu'au signe correspond un sens déterminé et au sens une référence déterminée tandis qu'à une référence (un objet) ne correspond pas un signe unique. De plus, un même sens a dans des langues différentes, et parfois dans la même langue, plusieurs expressions. À vrai dire, ce rapport régulier admet des exceptions. Dans un système de signes parfait, un sens déterminé devrait correspondre à chaque expression. Mais les langues vulgaires sont loin de satisfaire à cette exigence et l'on doit s'estimer heureux si dans le même texte, le même mot a toujours le même sens. » [Frege, 1971 : 104]

Cette dernière phrase témoigne du désintérêt que Frege porte à l'analyse linguistique et aux langues (en anglais le terme est traduit de l'allemand par *natural*). L'entreprise de l'auteur est en effet tout autre : construire un langage de formalisation logique. Pourquoi et dans quelles circonstances un objet pourrait-il être désigné par plusieurs expressions et plusieurs sens ? Frege ne nous en offre que le constat. Nous serions donc tenté de dire que le sens linguistique n'intéresse pas le logicien ; c'est la dénotation, parce qu'elle permet de déterminer la vérité des énoncés (on peut l'observer), qui est primordiale :

> « Mais pourquoi voulons-nous que tout nom propre ait une dénotation, en plus d'un sens ? Pourquoi la pensée ne nous suffit-elle pas ? C'est dans l'exacte mesure où nous importe sa valeur de vérité. Et tel n'est pas toujours le cas. Si l'on écoute une épopée, outre les belles sonorités de la langue, seuls le sens des propositions et les représentations ou sentiments que ce sens éveille tiennent l'attention captive. À vouloir chercher la vérité, on délaisserait le plaisir artistique pour l'examen scientifique. » [*ibid.* : 109]

En d'autres termes, la poésie, la fiction n'ont aucun intérêt du point de vue de la vérité : les événements qui y sont relatés n'existent pas puisqu'ils ne sont pas observables. Cette définition du sens, on l'aura compris, quand bien même elle prend en compte l'objet dans la définition du sens, est insuffisante du point de vue d'une sémantique linguistique.

Une des œuvres majeures sur la référence est la théorie référentielle de la communication de C. Ogden et I. Richards [Ogden & Richards, 1923], qui prend le contre-pied de la logique de Frege : cette théorie considère la référence comme un phénomène psychologique. Leur théorie est souvent évoquée par la reproduction du triangle sémiotique, illustré dans la figure (1.4). Les auteurs critiquent explicitement de Saussure pour avoir négligé les choses que les signes représentent [*ibid.* : 6] et proposent de les (ré-)introduire.

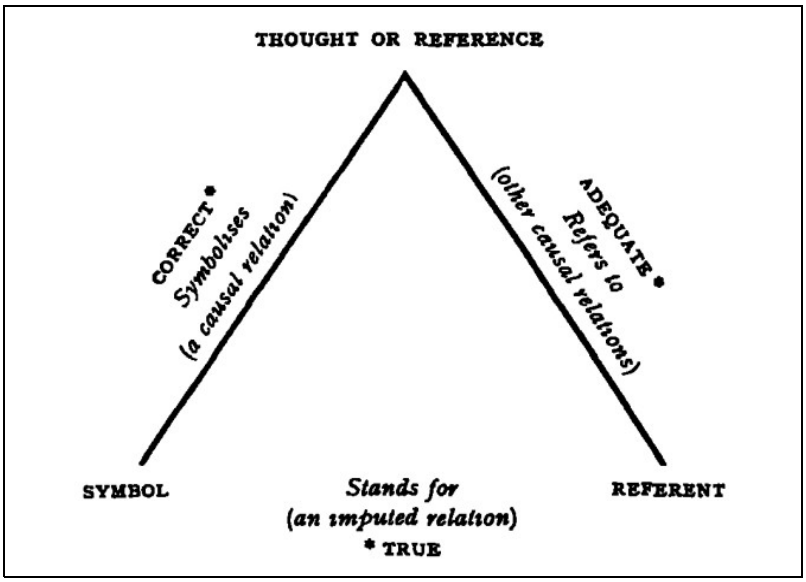


FIG 1.4 – « Le triangle sémiotique » (Ogden & Richards, 1923 : 11)

Pour comprendre ce triangle, il faut savoir qu'il veut restituer la dynamique de tout acte de communication. On peut en effet situer le locuteur et l'interlocuteur au centre (« Thought or Reference »). Les autres éléments du triangle (« Symbol » et « Referent ») sont extérieurs, mais doivent leur existence à la référence. Le schéma traduit donc l'idée suivante : il n'y a d'autre relation entre un symbole (mot en usage) et un référent que celle qu'effectue un locuteur/interlocuteur. Les auteurs s'attaquent en effet à cette illusion qui consiste à associer des objets ou des sens à des mots : les mots ne veulent rien dire en soi, pas plus que les objets. Il n'y a pas de sens sans interprétation :

> « It is only when a thinker makes use of [words] that they stand for anything, or, in one sense, have 'meaning.' » [*ibid.* : 10]

Les auteurs consacrent un chapitre entier pour répertorier les pouvoirs magiques qui ont pu être attribués aux mots [*ibid.* : ch. 2]. Ils ajoutent que les mots ont d'autres usages que la référence, comme de provoquer des émotions. Un mot devient le symbole d'une chose lorsqu'il est pensé par un individu comme tel. La question primordiale pour les auteurs est de savoir comment on réfère, en d'autres termes, d'expliciter le processus d'interprétation d'un signe. Leur théorie est que les mots deviennent associés aux choses par l'expérience répétée de leur association. Ce modèle, d'inspiration béhavioriste, définit la pensée comme l'enregistrement de ces associations enregistrées (les « engrams ») :

> « This simple case is typical of all interpretation, the peculiarity of interpretation being that when a context has affected us in the past the recurrence of merely a part of the context will cause us to react in the way in which we reacted before. A sign is always a stimulus similar to some part of an original stimulus and sufficient to call up the engram formed by that stimulus.
> An engram is the residual trace of an adaptation made by the organism to a stimulus. The mental process due to the calling up of an engram is a similar adaptation: so far as it is cognitive, what it is adapted to is its referent, and is what the sign which excites it stands for or signifies. » [*ibid.* : 53]

Le sens d'un mot (mais il en va de même de tout autre symbole) s'élabore et s'adapte en contexte, en tenant compte du contexte psychologique et externe :

> « To speak of a reference is to speak of the contexts psychological and external by which a sign is linked to its referent. » [*ibid.* : 68]

Pour se comprendre, un locuteur et un interlocuteur doivent par conséquent effectuer des actes de référence similaires, et, qui plus est, avoir partagé des expériences similaires.

> « Thus a symbol becomes when uttered, in virtue of being so caused, a sign to a hearer of an act of reference. But this act, except where difficulty in understanding occurs, is of little interest in itself, and the symbol is usually taken as a sign of what it stands for, namely that to which the reference which it symbolizes refers. When this interpretation is succesful it follows that the hearer makes a reference similar in all relevant respects to that made by the speaker. » [*ibid.* : 205]

On pourrait croire de notre description du triangle sémiotique, qu'il n'y a pas de place dans cette théorie pour une analyse linguistique des mots, puisque leur sens est leur référence et que la référence est un ensemble d'opérations mentales. Bien au contraire, le sens est symbolisé, c'est un mot en contexte ; le sens, en tant que symbole, peut tout à fait faire l'objet d'une théorisation, d'une

conceptualisation, tant qu'on garde à l'esprit qu'il n'est que symbole.

> « In discussion of method or of mental processes, 'concepts' or abstract references may, of course, be themselves talked about; and in this special case words will properly be said to stand for ideas. […] Words, as we have seen, always *symbolize* (cf. p. 11) thoughts and the conceptualist is apt to imply that the very special case of the construct or concept imagined for the purpose of an attempted scientific reference or classification, and then itself examined, can be generalized. He then states that the word is not a mere word as the nominalist holds, but stands for a conceptual symbol. » [*ibid.* : 100]

De fait, lorsqu'on interprète la référence au centre du triangle comme un concept, intermédiaire entre mot et objet, on dénature leur théorie, puisque les concepts sont des symboles. La référence de Ogden et Richards peut être assimilée à une tâche de fond qui opère dans la compréhension de symboles et nous permet de savoir que l'on franchit un grand pas lorsqu'on parle de sens d'un mot. À cet égard, les systèmes symboliques confectionnés par les linguistes (pour ce qui nous intéresse ici) sont en quelque sortes des langages de haut niveau, pour prendre une métaphore informatique. Pour reprendre le vocabulaire des auteurs, les linguistes ont la délicate tâche de décrire au moyen de symboles, les symboles (mots) employés au quotidien par quantité de locuteurs. Pour décrire ces premiers, nous utiliserons le terme de catégories linguistiques. Or, un certain nombre de catégories permettent d'établir une référence vers le réel, et c'est vers celles-ci que nous nous tournons à présent.

1.2.2. Aspects référentiels des catégories linguistiques

L'existence d'un certain nombre d'unités linguistiques semble justifier l'ancrage du langage dans le réel : elle ne peuvent s'interpréter autrement que vis-à-vis d'une situation d'énonciation dans laquelle des locuteurs interagissent avec le monde qui les entoure. On peut ainsi songer aux démonstratifs, dont la fonction principale est de « montrer » des choses (« monstration ») : *ceci, cela, ça*, par exemple, peuvent s'employer pour désigner concrètement des objets dans des situations réelles (référence dite « exophorique » ; ils peuvent également avoir une fonction anaphorique). Les pronoms personnels (de première et de seconde personne) permettent aux locuteurs de se représenter comme participants à des événements, et plus généralement, de s'identifier mutuellement à travers le langage. Un pronom comme *je* peut être employé par une multitude de locuteurs mais ne désignera à chaque fois que celui qui le prononce. Ces unités linguistiques ont fait l'objet d'une théorisation qui porte le nom de « deixis » [Bühler, 2009], reprenant la métaphore du repère à plusieurs dimensions, dans lequel tout acte de parole se situe. Les dimensions principales étant les interlocuteurs (*je, tu, mon, ton, moi, toi*, ...), l'espace (*ici, là*) et le temps (*aujourd'hui, hier,* ...). Si la deixis n'est pas l'unique fonction langagière, elle nécessite néanmoins d'être prise en compte et l'on voit mal comment réfuter le fait que le langage est tourné vers le réel.

Un autre type d'unité linguistique, qui peut passer inaperçu pour les linguistes, mais qui n'en a pas moins fait couler beaucoup d'encre (en logique notamment) est la catégorie de « nom propre ». Intuitivement, un nom propre permet de faire référence à une entité particulière, qui, d'usage, n'est pas nécessairement présente dans la situation d'énonciation (à la différence de déictiques stricts) : ce nom propre implique que l'entité désignée, le référent, existe, a existé ou existera, et, en tant qu'il sert à désigner un individu unique, n'a pas (généralement ; voir néanmoins le procédé rhétorique d'antonomase ainsi que les travaux sur les syntagmes à nom propre [Noailly, 1995]) un « sens » qui puisse s'appliquer à d'autres entités. À l'inverse du nom commun, il ne désigne pas une catégorie

et/ou une classe de choses, car il ne peut s'appliquer qu'à un individu. Il est certes possible de considérer que ce ne sont pas des unités linguistiques (des signes), comme le nom *livre* par exemple, et qu'à ce titre, ils ne figurent pas parmi les objets d'analyse linguistique : cela signifierait que le linguiste ne s'intéresse pas à tous les mots qui forment un texte, mais seulement à une sous-catégorie de mots (lesquels ? Sur quels critères ?). Nous reviendrons sur les différentes conceptions du Nom Propre et sur les difficultés liées à son application en TAL à travers la notion d'Entité Nommée (chapitre 5).

Si le paradigme référentiel a pu être réduit à une conception nomenclaturiste de la langue, c'est tout d'abord parce qu'il donne une place privilégiée au nom. La catégorie de nom (par opposition au verbe par exemple) renvoie aux notions de « (dé-)nomination », de « désignation » et de « référence » : les noms, comme *chat, livre, tasse*, nommeraient des classes de choses qui « existent » dans la réalité et permettraient de distinguer, d'organiser des types de choses. La détermination du nom peut permettre d'accomplir un acte de référence, et ceci, de plusieurs manières. En introduisant un nom par un pronom démonstratif (*ce chat*), ou en combinant un pronom défini à un nom et une relative restrictive (*le chat que j'ai vu hier*), on crée une expression dont le référent est unique et identifiable (voir [Charolles, 2002] pour une présentation générale des expressions référentielles).

On ne peut manquer de contraster les expressions démonstratives (1) avec les expressions définies (2) et indéfinies (3) dans le cadre de la grammaire du syntagme nominal.

(1)Donne-moi *cette* orange.
(2)Donne-moi *l'*orange.
(3)Donne-moi *une* orange.

Les exemples (1) et (2) s'opposent à l'exemple (3). La lecture de ce dernier est dite « indéfinie » car les différents référents possibles sont indistinctement considérés (glose : *une, peu importe laquelle*), alors que la lecture des deux premiers est définie. Dans les trois cas, la lecture est spécifique (non-générique, elle ne réfère pas à la classe) et la différence entre les deux premiers est subtile : dans le second, il n'y a pas d'ambiguïté sur le référent (peut-être ne reste-t-il qu'une orange), alors que le démonstratif implique qu'il peut y en avoir une (*cette orange et non celle-là*). Mais il est également évident que le sens du nom contribue à la réussite de cet acte de référence. Ainsi, pour prendre un exemple simple, la substitution du nom *orange* par *pomme* (4) implique un changement de référent.

(4)Donne-moi une *pomme*.

La référence suppose donc que la signification du nom « corresponde » au référent sous peine d'incompréhension. On peut également constater qu'un référent peut être désigné par plusieurs noms parce qu'il appartient à plusieurs classes. Ainsi, en prononçant l'énoncé de l'exemple (5), on fait référence à un objet qui appartient à une classe plus vaste que celle de *pomme* ou d'*orange*.

(5)Donne-moi le *fruit*.

Mais si le langage permet effectivement, à travers l'entremise des noms, de classer des objets, l'identification des similarités des objets ainsi classés et des critères de classement reste un domaine ouvert. L'objet de cette recherche porte aujourd'hui le nom de Catégorisation [Rosch et al., 1976] [Kleiber, 1990]. Pour aborder le problème, nous proposons de remonter le temps et de décrire le modèle de J. S. Mill, dans lequel les noms sont classés par rapport à des critères référentiels.

1.2.3. Le modèle de J. S. Mill

Un des objectifs de *System of Logic* de J. S. Mill [Mill, 1884] était de proposer des méthodes inductives (cinq) pour l'investigation scientifique. Ces méthodes devaient permettre de pouvoir obtenir des certitudes à partir de la généralisation d'observations empiriques (inférence). Le système de logique qu'il défend s'appuie sur une analyse du langage et plus particulièrement des noms généraux (« general names ») qui repose sur une double propriété : la capacité à signifier des attributs et celle à référer à une classe d'objets. Pour distinguer l'usage anglais *name*, par opposition à *noun*, tous deux traduits par « nom » en français, nous utiliserons le terme *dénomination* pour *names* (plutôt que *nom-names* comme le fait Kleiber).

1.2.3.1. Les dénominations comme classes de choses

Mill propose de distinguer les dénominations en fonction de la connotation :

- le terme non-connotatif peut renvoyer à un sujet (« une chose ») ou à un attribut (« une classe de choses » ; dite dénomination abstraite) qui n'ont d'autres sens que ce qu'ils dénotent. Le nom propre *John* est une dénomination non-connotative car il « dénote » uniquement un sujet.
- la dénomination connotative (ou dénomination générale ; « general name ») possède deux dimensions de sens, il dénote la classe de sujets et il connote (implique) un (ou plusieurs) attribut. L'adjectif *blanc* est une dénomination connotative car il dénote *toutes les choses blanches* et connote l'attribut « blancheur ».

Chaque dénomination peut être associée à une classe de choses. Pour Mill, le sens réside dans la connotation, mais celui-ci est déduit (au sens empirique) de notre contact avec le monde. À l'inverse des noms propres qui désignent directement des individus en question, les dénominations comme *homme* ne les désignent qu'indirectement, via la prédication de propriétés :

> « All concrete general names are connotative. The word man, for example, denotes Peter, Jane, John, and an indefinite number of other individuals, of whom, taken as a class, it is the name. But it is applied to them, because they possess, and to signify that they possess, certain attributes. There seem to be corporeity, animal life, rationality, and a certain external form, which for distinction we call the human. Every existing thing, which possessed all these attributes, would be called a man; and any thing which possessed none of them, or only one, or two, or even three of them without the fourth, would not be so called. For example, if in the interior of Africa there were to be discovered a race of animals possessing reason equal to that of human beings, but with the form of an elephant, they would not be called men. Swift's Houyhnhnms would not be so called. Or if such newly-discovered beings possessed the form of man without any vestige of reason, it is probable that some other name than that of man would be found for them. How it happens that there can be any doubt about the matter, will appear hereafter. The word man, therefore, signifies all these attributes and all subjects which possess these attributes. But it can be predicated only of the subjects. What we call men, are the subjects, the individual Stiles and Nokes; not the qualities by which their humanity is constituted. The name, therefore, is said to signify the subjects directly, the attributes indirectly; it denotes the subjects, and implies, or involves, or indicates, or, as we shall say henceforth connotes, the attributes. » [*ibid.* : 35]

La définition des dénominations, chez Mill, est une analyse, c'est-à-dire une décomposition en fonction d'attributs (ou de traits sémantiques) ; les attributs connotés par des dénominations pouvant eux-mêmes être décomposés en attributs, puis, à terme, en les parties des phénomènes indivisibles qui les ancrent dans le réel [*ibid.* : 107] : pour l'attribut *blancheur*, il s'agit de « la propriété ou le pouvoir d'exciter la sensation de blanc » [*ibid.*], la « sensation de blanc » ne pouvant être définie qu'en faisant appel à l'expérience personnelle de l'interlocuteur [*ibid.*]. Le langage nous permet ainsi de classer les choses :

> « There is, as has been frequently remarked in this work, a classification of things, which is inseparable from the fact of giving them general names. Every name which connotes an attribute, divides, by that very fact, all things whatever into two classes, those which have the attribute and those which it can not. And the division thus made is not merely a division of such things as actually exist, or are known to exist, but of all such as may hereafter be discovered, and even of all which can be imagined. » [*ibid.* : 497]

Mill, comme on peut s'en rendre compte, ne limite pas la classification des choses à ce qui existe concrètement, mais à tout ce qui peut être perçu, ressenti, imaginé et pensé.

1.2.3.2. Dénomination et Classification

Les choses seraient simples si le langage n'était qu'une nomenclature, où chaque segment de la réalité porteraient un nom et où ces noms nous permettraient de désigner sans ambiguïté ces segments. Il n'en est rien. Il faut donc distinguer les catégories linguistiques et les classes référentielles pour mieux analyser leurs interactions. Comme le rappelle Mill, si le langage permet d'imposer sur la réalité des catégories de sens, ces dernières peuvent trouver leur justification par la perception de similarités entre référents (dans leur fonctionnement, leur forme, etc.) :

> « By every general name which we introduce, we create a class, if there be any things, real or imaginary, to compose it; that is, any Things corresponding to the signification of the name. Classes, therefore, mostly owe their existence to general language. But general language, also, though that is not the most common case, sometimes owes its existence to classes. A general, which is as much to say a significant, name, is indeed mostly introduced because we have a signification to express by it; because we need a word by means of which to predicate the attributes which it connotes. But it is also true that a name is sometimes introduced because we have found it convenient to create a class; because we have thought it useful for the regulation of our mental operations, that a certain group of objects should be thought of together. » [*ibid.* : 94]

Mill introduit ici une idée importante sur la question des relations entre la réalité (classes) et le langage : le langage permet de classer les choses, ainsi que de nommer des classes de choses ; il est à la fois influencé par la réalité et il la structure.

L'analyse des catégories (dénominations) ou classes référentielles porte le nom de classification ou de catégorisation. Cette opération peut être conçue de manière ascendante, en regroupant en priorité les référents les plus similaires, ou de manière descendante, en divisant les ensembles de référents selon leur degré de dissimilitude. La construction d'une classification suppose que l'on s'interroge sur la pertinence des critères d'agrégation de classes : leurs attributs. De plus, la classification ne se limite pas à classer des individus, certaines catégories ont plus d'importance que d'autres.

La tradition aristotélicienne distingue deux types de relations catégorielles majeures, les Genres et les Espèces. Une classe est dite espèce, ou genre, respectivement à une autre. Ces catégories ne définissent pas la classe en soi, mais qualifient la relation entre deux classes, l'une (le genre) incluant l'autre (l'espèce). Par exemple, la même dénomination générale *animal* est un « genre » vis-à-vis d'*homme*, mais une espèce vis-à-vis d'*être vivant.* Comme le rappelle Mill, l'espèce doit nécessairement posséder les attributs du genre (pour être dite espèce de ce genre), ainsi que des attributs spécifiques (« Differentia ») qui le distinguent d'autres espèces du même genre : on connaît ce processus sous la dénomination de *contrainte d'héritage*, qui permet d'établir des hiérarchies de structures genre-espèce de la plus générique à la plus spécifique, comme les taxonomies et les ontologies.

D'après Mill, il faut distinguer les classes établies sur des critères fortuits de celles qui rendent compte d'une véritable différence. Les aristotéliciens formulaient cette différence sur la base de « l'essence », ou cause suprême sans laquelle une chose ou une classe de choses ne saurait exister. Mill critique cette vision simpliste sans pour autant rejeter la distinction. Les catégories existent mais elles se caractérisent d'après l'inhexaustibilité de ses attributs, donnant ainsi un moyen de tester cette « catégorialité » :

> « It appears, therefore, that the properties, on which we ground our classes, sometimes exhaust all that the class has in common, or contain it all by some mode of implication; but in other instances we make a selection of a few properties from among not only a greater number, but a number inexhaustible by us, and to which as we know no bounds, they may, so far as we are concerned, be regarded as infinite. » [*ibid.* : 97]

Les catégories (référentielles) sont celles qui regroupent des individus ayant un nombre « infini » de points communs. Par déduction, une dénomination générale désigne une catégorie dans la mesure où elle implique un nombre infini d'attributs partagés par les membres de la classe correspondante. D'après Mill, l'erreur des aristotéliciens est d'avoir recherché l'attribut « essentiel » (Differentia) qui explique cette distinction respectivement à chaque classe, en reléguant les autres au statut d'attribut contingent (Accidens). C'est donc dans la définition, dans la connotation, dans les dénominations (et non dans les choses elles-mêmes), que s'effectue cette distinction qui permet d'obtenir des catégories.

1.2.3.3. Dénomination et sens

Dans la logique de Mill, la Classification affronte nécessairement la question de la variation du sens de la dénomination [*ibid.* : 78, 480]. Par exemple, une dénomination pourra avoir différents sens selon l'objectif de la classification, car on peut concevoir un même objet selon différents points de vue [*ibid.* : 101]. Mill se refuse à imposer une classification fixe des choses qui est contraire à toute approche empirique. Sa logique s'accompagne d'une conception relativiste de la connaissance. Les connaissances évoluent, les dénominations d'hier ne sont pas nécessairement celles d'aujourd'hui, les classes dénotées par une dénomination varient. Ce qui est remarquable, sous cet angle, c'est qu'une dénomination non-connotative peut être considérée comme connotative d'un certain point de vue, et pourquoi pas, dans le cadre d'une certaine pratique sociale. Les noms propres et les attributs peuvent à leur tour être redéfinis s'ils remplissent les conditions :

> « Words not otherwise connotative may, in the mode just adverted to, acquire a special or technical connotation. Thus the word whiteness, as we have so often remarked, connotes nothing; it merely denotes the attribute corresponding to a certain sensation; but if we are making a classification of colors, and desire to justify, or even merely to

> point out, the particular place assigned to whiteness in our arrangement, we may define it "the color produced by the mixture of all the simple rays;" » [*ibid.* : 102]
> « 'and the same objects, therefore, may admit with propriety of several different classifications. Each science or art forms its classification of things according to which fall within its special cognizance, or of which it must take account in order to accomplish its peculiar practical end. A farmer does not divide plants, like a botanist, into dicotyledonous and monocotyledonous, but into useful plants and weeds. » [*ibid.* : 500]

L'auteur note également que le contexte (les circonstances) peut contraindre l'usage des mots, de sorte à ce que leur sens inclue ces contraintes :

> « It continually happens that of two words, whose dictionary meanings are either the same or very slightly different, one will be the proper word to use in one set of circumstances, another in another, without its being possible to show how the custom of so employing them originally grew up. The accident that one of the words was used and not the other on a particular social circle, will be sufficient to produce so strong an association between the word and some speciality of circumstances, that mankind abandon the use of it in any other case, and the speciality becomes part of its signification. » [*ibid.* : 481]

L'auteur donne l'exemple de *loyalty* qui était au départ restreint à une fidélité à l'égard du trône. La signification d'un mot peut donc également intégrer (« becomes part ») les « spécialités des circonstances », ou contexte spécifique (qu'il reste à caractériser) dans lequel il est employé. Une dénomination, seule, décontextualisée, ne peut nous apprendre grand chose, il faut rechercher dans le contexte les éléments qui participent de sa connotation. Définir les dénominations, c'est les comprendre, c'est permettre de connaître un peu mieux les objets qu'ils dénotent. En prenant en compte la totalité des usages d'une dénomination, on élargit ainsi notre champ de connaissance :

> « We thus see that to frame a good definition of a name already in use, is not a matter of choice but of discussion, and discussion not merely respecting the usage of language, but respecting the properties of things, and even the origin of those properties. And hence every enlargement of our knowledge of the objects to which the name is applied, is liable to suggest an improvement in the definition. »[*ibid.* : 470]

De plus, les connotations évoluent en fonction des objets auxquels elles s'appliquent. Mill cite à ce propos la loi de transition de Stewart[2], que l'on peut résumer par le terme de *polysémie diachronique*, selon laquelle l'usage des mots tend à s'écarter, par transitions historiques, de son sens de départ. Une dénomination qui ne s'appliquait qu'à une certaine catégorie, s'applique progressivement à d'autres par ressemblance entre deux catégories, au détriment de l'homogénéité de la connotation :

> « And (without attempting to decide a question which in no respect belongs to logic) I can not but feel, with him [Stewart], considerable doubt whether the word beautiful connotes the same property when we speak of a beautiful color, a beautiful face, a beautiful scene, a beautiful character, and a beautiful poem. The word was doubtless extended from one of these objects to another on account of a resemblance between

2 La loi de transition est le fondement de la théorie du prototype en ressemblance de famille [Kleiber, 1990] [Lakoff, 1987] [Wittgenstein, 1953].

> them, or, more probably, between the emotions they excited; and, by this progressive extension, it has at last reached things very remote from those objects of sight to which there is no doubt that it was first appropriated; » [Mill, 1884 : 475]

Chercher à associer tous ces usages à une seule définition (un invariant) serait possible, mais laisserait de côté une portion du sens que le nom peut véhiculer. Mill propose, dans ce genre de cas, de restreindre une définition en fonction des objets auxquels il s'applique :

> « It is better, in such a case, to give a fixed connotation to the term by restricting, than by extending its use; rather excluding from the epithet Beautiful some things which it is commonly considered applicable, than leaving out of its connotation any of the qualities by which, though occasionally lost sight of, the general mind may have been habitually guided in the commonest and most interesting applications of the term. » [*ibid.* : 475]

Cette proposition est intéressante car elle amène à chercher des contraintes de sens d'une dénomination dans les dénominations avec lesquelles elle se combinent (voir la notion de collocation, chapitre 2). Comme les dénominations impliquent des attributs chez Mill, cela signifierait que les associations d'attributs stabilisent le sens. C'est donc, semble-t-il, dans la grammaire des dénominations, que Mill nous invite à caractériser le sens des noms.

1.2.3.4. Point de vue de Mill sur le nom propre

La conception du langage que nous venons d'ébaucher n'est pas ce que les logiciens ont retenu de l'œuvre de Mill (voir néanmoins [Kleiber, 1981]). On retient généralement sa position sur les noms propres, qu'il considère dénués de sens (de connotation) :

> « Proper names are not connotative: they denote the individuals who are called by them; but they do not indicate or imply any attributes as belonging to those individuals. When we name a child by the name Paul, or a dog by the name Cæsar, these names are simply marks used to enable those individuals to be made subjects of discourse. It may be said, indeed, that we must have had some reason for giving them those names rather than any others: and this is true; but the name, once given, becomes independent of the reason. » [Mill, 1884 : 36]
> « From the preceding observations it will easily be collected, that whenever the names given to objects convey any information, that is, whenever they have properly any meaning, the meaning resides not in what they *denote*, but in what they *connote*. The only names of objects which connote nothing are *proper* names; and these have, strictly speaking, no signification. » [*ibid.* : 37]

Partant, un nom propre ne peut être défini.

> « The definition of a word being the proposition which enunciates its meaning, words which have no meaning are unsusceptible of definition. Proper names, therefore, can not be defined. A proper name being a mere mark put upon an individual, and of which it is the characteristic property to be destitute of meaning, its meaning can not of course be declared; though we may indicate by language, as we might indicate still more conveniently by pointing with the finger, upon what individual that particular mark has been, or is intended to be, put. It is no definition of « John Thomson » to say he is « the son of General Thomson; » for the name John Thomson does not express this. Neither is

> it any definition of « John Thomson » to say he is « the man now crossing the street. » These propositions may serve to make known who is the particular man to whom the name belongs, but that may be done still more unambiguously by pointing to him, which, however, has not been esteemed one of the modes of definition. » [*ibid.* : 105]

Le problème qu'on rencontre est de pouvoir distinguer le sens des phrases contenant un même référent, lorsque celui-ci est ou n'est pas désigné par le même nom propre (6-7).

(6)Romain Gary est Émile Ajar.
(7)Romain Gary est Romain Gary.

Les phrases (6) et (7) ont effectivement un sens différent pour qui est familier de ces noms propres, l'une révèle la réelle identité d'un individu, l'autre ressemble à une tautologie. Or, si le sens d'un nom propre réside *uniquement* en sa dénotation, on ne peut capter cette subtilité sémantique. Notons néanmoins qu'avant l'assertion de la phrase (6), les deux noms propres pouvaient être conçus comme deux référents distincts (l'affaire est en fait plus complexe lorsque l'on étudie la vie de cet écrivain) et que la phrase (7) peut avoir une fonction pragmatique différente selon le contexte (son sens ne serait pas uniquement redondant). Pour Mill, ce type d'exemple est rare et signifie que ces deux noms dénotent le même objet :

> « The only propositions of which Hobbes' principle is a sufficient account, are that limited and unimportant class in which both the predicate and the subject are proper names. For, as has already been remarked, proper names have strictly no meaning; they are mere marks for individual objects: and when a proper name is predicated of another proper name, all the signification conveyed is, that both the names are marks for the same object. » [*ibid.* : 101]

Une autre difficulté concerne le fait que certaines dénominations (comme *le soleil*, *la lune*), qu'on appelle aussi des unica, ont un sens mais ne désignent qu'un individu.

> « But there is another kind of names, which although they are individual names, that is, predicable only of one object, are really connotative. For, although we may give to an individual a name utterly unmeaning, which we call a proper name,—a word which answers the purpose of showing what thing it is we are talking about, but not of telling anything about it; yet a name peculiar to an individual is not necessarily of this description. It may be significant of some attribute, or some union of attributes, which being possessed by no object but one, determines the name exclusively to that individual. “The sun” is a name of this description; » [*ibid.* : 34]

Pour Mill, ces dénominations sont générales, car rien dans leur connotation n'empêche d'en imaginer plusieurs [*ibid.*]. Il considère également que des dénominations comme *l'armée de César*, lorsqu'on fait référence à une bataille particulière par exemple, ne dénotent qu'un individu. Dans ce cas ce sont des dénominations générales *employées* comme dénominations individuelles (elles remplissent la fonction de noms propres). Ces observations portent à croire que le critère d'individualité n'est pas définitoire pour les noms propres. Il faut donc distinguer deux niveaux, qui ne coïncident pas toujours : dénotation-connotation et individu-général.

1.2.4. La mise au point de Kleiber

L'analyse de la référence suppose qu'on élargisse le champ des choses que l'on peut désigner à celles qui n'existent pas dans la réalité, comme les idées, les entités mathématiques, ou les êtres surnaturels (par exemple 1.2.3.1). Il en est de même de ce que l'on peut connaître empiriquement et de ce que l'on ne peut connaître que par médiation : il nous semble naturel d'accorder une existence à des planètes ou même des espèces sous-marines dont on nous parle sans pour autant les avoir vues, connues par nous-mêmes. Tous les objets auxquels on peut référer n'ont pas le même statut d'existence ni de connaissance.

On sait également que le « monde » auquel nous référons ne peut être connu objectivement : notre perception du monde dépend de nos capacités biologiques, perceptuelles et psychologiques et varie de ce point de vue de la perception d'autres espèces. De plus, comme le montrent les travaux de la théorie de la Gestalt [Koffka, 1935], notre monde existe dans une certaine forme, déjà catégorisé, interprété par notre perception.

G. Kleiber note à juste titre que si, effectivement, le monde est une réalité perçue, alors les noms ne sont plus des noms de choses, mais de choses perçues [Kleiber, 1999a : 20]. Ce fait n'empêche pas de parler du réel « comme si c'était... le réel » [*ibid.* : 22, 27] : l'ex-réalité objective peut être redéfinie comme réalité inter-subjective, correspondant à ce que l'on *appelle* la réalité objective, illustrée par nos capacités à nous entendre, à nous comprendre, qui comme nous l'avons vu, ne peut être remise en cause (cf. *infra* 1.1.2). Ce qui compte, c'est ce « jeu-de-langage » [Wittgenstein, 1953] dont une des règles est la supposition de l'indépendance de la réalité :

> « Que cette entité n'ait pas d'existence objective, indépendante, en somme, que son existence soit due à l'interaction homme-réalité, importe peu : nous l'appréhendons – mythe ou non – comme s'il en allait ainsi, comme si c'était une portion de réalité, indépendante du langage. » [Kleiber, 1999a : 26]
> « Et c'est précisément parce que nous croyons qu'il existe un monde réel avec des individus et des choses « réels » que la référence à des mondes et à des individus et des choses non réels est envisageable. Autrement dit, si la réalité n'était pas ce que nous pensons qu'elle est, à savoir réelle, nous ne pourrions pas concevoir des entités et choses fictives ou imaginaires et nous ne pourrions donc pas référer au Père Noël, aux licornes et à plein d'autres choses encore. La notion de monde possible n'a de sens que par rapport à un monde réel, qui possède un statut privilégié. Le potentiel et l'irréel ou contrefactuel présupposent le réel. » [*ibid.* : 23]

Cette argumentation posée, Kleiber s'attache ensuite à analyser le statut de la référence dans la sémantique. Son argument principal est que si l'on accepte que le langage permet de parler du réel, alors le sens doit, d'une manière ou d'une autre, refléter cet aspect : il serait ainsi étrange de définir deux dimensions indépendantes du sens, l'une référentielle, l'autre conceptuelle (par exemple), puisque l'une pourrait se passer de l'autre. Son argumentaire a pour objectif de montrer que toute théorie du sens repose en dernier lieu sur la référence. Il propose une vision hétérogène du sens, toujours de nature référentielle, dans laquelle certaines dimensions sont plus appropriées que d'autres à la description de phénomènes linguistiques :

> « L'hypothèse que nous suggérons est que le sens obéit à deux modèles référentiels différents : le modèle descriptif, celui qui indique quelles sont les conditions (nécessaires et suffisantes ou prototypiques) auxquelles doit satisfaire une entité pour pouvoir être désignée ainsi, et le modèle instructionnel, qui marque le moyen d'accéder au ou de construire le référent. […] Les conceptions aréférentielles du sens que nous

> venons de présenter commettent, en fait, la même erreur qu'a commise la théorie combattue : celle de croire que le sens est homogène et donc qu'il ne peut être que d'une seule nature : ou référentiel ou instructionnel (argumentatif ou non), ou abstracto-dynamique, etc. […] il convient de prôner un sens hétérogène, qui peut varier selon le type d'expression […], accepter qu'une même unité puisse présenter du sens mixte, relevant du statut descriptif et du statut instructionnel. » [*ibid.* : 50]

Kleiber s'inscrit dans la lignée des travaux de Mill, l'une des différences majeures étant qu'il substitue à la notion de dénotation chez Mill le sens instructionnel (qui, comme pour les déictiques, sont des indications permettant d'identifier le référent en contexte) pour le cas du Nom Propre, et le sens descriptif à la notion de connotation chez Mill.

1.2.5. Polysémie et Métonymie intégrée

Le problème de la nature du sens semble partiellement résolu par l'invocation du paradigme référentiel. Ce dernier repose sur la dualité classe / catégorie, la première correspondant à l'extension des référents auxquels une expression linguistique s'applique, la seconde à l'ensemble des propriétés sémantiques qu'ils partagent. Néanmoins, il reste à aborder un second problème qui est celui de la polysémie (voir également 1.3.2). La polysémie tient en deux choses [Cadiot & Habert eds., 1997] : la pluralité de sens d'une même forme et la perception d'un lien entre ces sens. Elle exige également la prise en compte du contexte. Si donc, une unité linguistique désigne une classe de choses, une catégorie référentielle, comment expliquer les variations de sens que l'on perçoit en (8)-(9) et en (10)-(11) :

(8) Je me suis fait *voler* mon appareil photo !
(9) L'avion *vole* au-dessus de l'océan
(10) Il m'a jeté le *livre* à la figure.
(11) Ce *livre* aborde des problèmes de sémantique.

Les occurrences de *voler* dans les exemples (8) et (9) sont généralement considérées (en lexicographie, par exemple) comme des verbes homonymes : il n'y a aucun rapport de sens entre le fait pour une entité de se déplacer dans les airs et celui de s'approprier un bien d'autrui sans son consentement. Ces deux verbes correspondent donc à deux signes (dénominations, symboles, etc.) différents.
Pourrait-on en dire autant du nom *livre* dans les contextes des exemples (10) et (11) ? S'il n'y a aucun rapport entre *jeter* et *aborder*, il demeure qu'un livre est à la fois un objet concret et un support d'information. Doit-on considérer qu'il s'agit là de deux sens distincts, de deux objets distincts ? Si oui, comment rendre compte du lien qui unit ces deux interprétations ?

Si l'on souscrit à la thèse référentielle, il faut pouvoir expliquer ces variations à partir de la nature des référents que ces expressions linguistiques désignent. Différentes théories de la polysémie ont été proposées dans ce cadre [Cruse, 1986] [Lakoff, 1987] [Nunberg, 1978] [Pustejovsky, 1998], mais nous retiendrons ici le traitement de Kleiber [Kleiber, 1999b] [Kleiber, 1999a], au moyen du principe de métonymie intégrée.

Pour Kleiber, les exemples (10) et (11) désignent le même objet, le livre, il n'y a pas de transfert sémantique de l'entité (ou des entités) en tant qu'objet, à l'entité en tant que support d'information, ni non plus deux sens à *livre*. Ce n'est pas qu'une partie de l'objet qui est mis en exergue par les prédicats de *jeter* ou de *parler*, mais la totalité de l'objet qui est en cause. La sensation de variation sémantique qui peut apparaître en (10 - 11) est uniquement due au prédicat :

> « Un prédicat, comme nous l'avons affirmé à plusieurs reprises, peut être dit vrai d'une entité individuelle ou d'un ensemble d'individus sans que nécessairement toutes les parties ou tous les membres satisfassent à ce prédicat. Une « partie » d'un référent singulier ou collectif permet, dans des conditions que nous préciserons tout de suite, une assertion sur le référent tout entier (dans sa globalité), grâce à ce que nous avons appelé le *principe de métonymie intégrée* :
> 31) *Certaines caractéristiques de certaines parties peuvent caractériser le tout*
> Ce qui autorise le passage de la partie au tout, c'est que les caractéristiques concernées soient d'une manière ou d'une autre également saillantes ou valides pour le tout. Autrement dit, qu'elles aient une répercussion sur le référent saisi dans sa globalité et que ce sont ces raisons qui font que c'est le référent global qui est choisi comme sujet et non seulement la partie vérifiant plus étroitement ou plus directement le prédicat. » [*ibid.* : 99]

Étant donné qu'on peut attribuer plusieurs parties à un même objet, il faut pouvoir expliquer pourquoi le choix se porte sur l'une d'entre elles. Kleiber pense qu'elles sont hiérarchisées par un principe de saillance : il propose une première définition de *livre* comme un « objet physique servant de support à un texte » [*ibid.* : note 21, 100], où la partie « objet concret » est la plus saillante. Il la compare à la définition de *roman*, dont l'aspect matériel est moins discriminant que l'aspect informationnel : « type de texte ayant pour support un objet appelé "livre" » [*ibid.*]. Cette différence expliquerait leur capacité à se combiner avec des prédicats « physiques » en (12) et (13).

(12) Un livre rouge/sale/déchiré [*ibid.* : 95]
(13) ? Un roman rouge/sale/déchiré

Ainsi, sans pour autant renier le fait qu'un roman est un livre, qu'ils possèdent donc des attributs (des parties) en commun, la différence majeure est que *roman* est de manière *saillante*, un type de texte, alors que *livre*, un type d'objet concret.

Ce principe (et les principes assimilés de méronimisation, aliénation et congruence ; [Kleiber, 1999b]) permet d'expliquer les phénomènes de métonymie à partir des relations entre les parties d'un objet et cet objet. Mais on rencontre sans cesse de nouveaux exemples qui tendent à fragiliser de tels principes. On connaît par exemple des livres blancs, des livres noirs qui ne sont pas dénommés à cause de la couleur de leur couverture. *Livre noir* exprime, par exemple, la composante « texte » de l'objet, c'est un livre de critique. *Livre blanc* désigne un état de l'art sur un sujet particulier. Dira-t-on que les noms de couleurs se combinent différemment avec la catégorie « livre » ? On pourrait également répondre que ces combinaisons /*livre+adjectif de couleur*/ n'ont pas tous le même statut : /*livre+noir*/ et /*livre+blanc*/ seraient, dans ces cas, des lexicalisations, des termes spécifiques faisant référence à une certaine catégorie de livre, dont la signification est plus figée que celle de /*livre+rouge*/ ; par conséquent, des livres dont la facette Texte est rendue plus saillante par ce phénomène de lexicalisation. Mais si tel était le cas, quels critères employer pour distinguer les deux interprétations lorsque l'on limite notre analyse au syntagme nominal (uniquement) ? De la même manière, *livre rouge* ne peut-il pas désigner les deux sens (14-15) ?

(14)Rappelons simplement qu'un livre rouge regroupe une ou plusieurs listes des noms de taxons les plus rares et menacés dans un territoire donné.[3]
(15) Ce livre rouge sur la flore européenne a également une couverture rouge.[4]

3 Source : http://www.naturecentre.org/biodiversite/67.html
4 Exemple inventé

Que dire des usages attestés de (16) ? C'est bien la composante matérielle qui est en jeu.

(16)En 1999, les Jeunesses poutiniennes ont déchiré son roman Le Lard Bleu en place publique.[5]

On s'aperçoit donc qu'une dimension supplémentaire entre en ligne de compte, le facteur idiomatique ou phraséologique : si l'on ne sait pas que *livre blanc* réfère à une certaine catégorie de livre pour lesquels la sensation de couleur sur la couverture n'a rien à voir avec l'usage qui est fait de l'adjectif (du moins synchroniquement) de couleur, l'interprétation se fera en termes compositionnels, c'est-à-dire en cherchant à associer aussi heureusement que possible la signification de *livre* à la signification de *blanc*. Une fois cette **connaissance** acquise, il faut pouvoir distinguer les deux usages de cette **construction**, du moins si l'on veut bien pousser l'analyse aussi loin. Pour cela il faut, en toute logique, « sortir du syntagme », et identifier des éléments dans le contexte. Deux voies, au moins, semblent possibles :

i. Ces deux syntagmes montreront des différences de position, de prédication verbale, d'appositions, disons, de dépendance syntagmatique (dans la phrase et au-delà, en prenant en compte les chaînes de co-référence par exemple), pouvant relever de l'idiomatique ou du compositionnel ;
ii. Ces deux syntagmes apparaitront dans des textes relevant de genres, de domaines, ou de thématiques différents : une interprétation sera plus saillante dans un contexte textuel que dans un autre.

L'analyse de textes peut se révéler pertinente pour dépasser certaines limites, proposer de nouvelles hypothèses dans le cadre du paradigme référentiel, les invalider et les tester. Cependant, cette entreprise est sans doute risquée, vu la quantité de textes que l'on peut étudier : comment sélectionner les textes, comment les regrouper et comment caractériser leurs similarités thématiques ou discursives ? L'ampleur de la difficulté des phénomènes textuels constituera-t-il un obstacle à l'identification de telles contraintes ?

5 Source : http://www.editions-verdier.fr/v3/oeuvre-roman.html

1.3. La tradition rhétorico-herméneutique

La tradition rhétorico-herméneutique s'inscrit dans le cadre de la sémiologie, l'analyse des signes, renouvelée dans la perspective saussurienne. À partir de la distinction langue/parole de Saussure, qui caractérisait la langue comme un *fait social*, cette tradition a évolué pour identifier les rapports qu'entretiennent les systèmes linguistiques et les activités sociales. La vision de cette tradition peut être résumée par le principe suivant :

> « L'usage d'une langue est par excellence une activité sociale, si bien que toute situation de communication est déterminée par une pratique sociale qui l'instaure et la contraint. » [Rastier, 1989 : 39]

Dans cette partie, nous rappellerons quelques aspects de la sémantique saussurienne avant d'aborder une présentation de la sémantique textuelle de F. Rastier.

1.3.1. Sémantique et Signe

Saussure, l'un des pères de la linguistique, adopte un point de vue ambigu vis-à-vis de la sémantique. Il considère que l'unité fondamentale, le signe, est le produit de l'association arbitraire entre une « image acoustique » et un concept. Il affirme que le sens est une représentation mentale (une unité de la pensée), ce qui suggère une vision conceptuelle du sens (un signe désigne une idée). En sus, il définit l'objet de la linguistique comme la description de la *langue*, conçue comme un système abstrait, fruit de conventions sociales partagées par la totalité des locuteurs : il adopte donc le point (iv) (p10). Pourtant, il n'analyse pas la nature ni la structure de cette représentation mentale : certains ont ainsi pu conclure que Saussure ne dit finalement rien sur la sémantique [Engler, 1973]. Une des explications est qu'il n'autorise pas d'autonomie au signifié vis-à-vis du signe : il n'adopte pas la thèse de l'autonomie de la sémantique. Comme le remarque Utaker, le sens est (dans) le signe, c'est-à-dire qu'on ne peut analyser le sens en linguistique qu'en faisant usage de l'unité (irréductible) qu'est le signe :

> « Le signe n'est pas le sensible ou une chose sensorielle qui renvoie à un sens ou à un concept d'une manière extrinsèque. Il est plutôt le lieu où le sensible (ou ce qui est sensoriel) est en même temps un sens. Au lieu d'une relation entre le signe et son référent, Saussure le pense par conséquent comme une unité qui fait que le sens est une propriété d'un signe qui par là a deux faces, le signifiant et le signifié. » [Utaker, 1996 : 51]

La sémantique que l'on peut entamer avec Saussure, est une sémantique du signe : séparer le signifiant du signifié serait un contre-sens. Mais la solidarité signifié-signifiant n'est pas suffisante. De fait, les associations entre signes existent et contribuent à sa caractérisation : elles s'établissent sur les plans syntagmatique (« la chaîne parlée ») et paradigmatique (associations de signes selon un critère donné). Mais ce sont des signes et non des signifiés (ou des idées, des choses) qui sont systématiquement mis en relation, même si le critère d'association repose sur une analogie de signifiés [de Saussure, 1916 : 174] : il y a autant d'associations possibles « qu'il y a de rapports divers » [*ibid.* : 173]. La totalité de ces rapports, le travail de caractérisation d'un signe par rapport aux autres, par opposition et différences, constitue ce que Saussure nomme la valeur (les valeurs ?) du signe.

Enfin, l'insistance (le dogmatisme, diraient Ogden et Richards) avec laquelle le principe d'arbitraire du signe a pu être défendu, masque la relativité que de Saussure lui attribuait. Le CLG (Cours de Linguistique Générale) est un cours et il est commun de débuter un cours par des principes généraux pour les relativiser à la fin. Le principe d'arbitraire du signe consiste à observer qu'il n'y a aucun lien entre le son et le sens dans la réalité, aucune justification pour leur association. En poursuivant cette logique, il semblerait qu'il ne soit pas possible que la structure de la langue, dans sa dimension paradigmatique et syntagmatique s'appuie sur des structures « extra-linguistiques ». Ce serait une lecture erronée du Cours, car l'arbitraire est une question de degré :

> « *le signe peut être relativement motivé*
> Ainsi *vingt* est immotivé, mais *dix-neuf* ne l'est pas au même degré, parce qu'il évoque les termes dont il se compose et d'autres qui lui sont associés, par exemple, *dix*, *neuf*, *vingt-neuf*, *dix-huit*, *soixante-dix,* etc. » [*ibid.* : 181]

Saussure formule l'opposition en pôles lexical et grammatical : la grammaire est le lieu de l'analyse des régularités de construction et peut révéler des phénomènes motivés.

1.3.2. De l'inutilité de la polysémie en sémantique

À présent que nous avons présenté le modèle saussurien, nous pouvons aborder le problème de la polysémie du signe chez de Saussure. Une des conséquences, qui peut paraître étrange, de sa conception du signe, veut que deux signifiés soient jugés différents parce qu'ils sont les propriétés de deux signes différents ; il n'y aurait, par voie de conséquence, pas de place pour l'existence de la polysémie, d'après le raisonnement suivant :

> « En ce sens une polysémie et une univocité sont seulement possibles en vertu de la séparation entre l'expression et le sens. Donc la polysémie pour Saussure ne peut pas exister étant donné que le même mot ou le même signe ne peut pas avoir plusieurs sens, tout comme deux mots ne peuvent pas exprimer le même sens (le cas corollaire de la polysémie ; la synonymie). L'unité du signe exclut à la fois la polysémie et la synonymie dans leur sens classique. Car ces termes impliquent une dissociation entre le signe et son sens qui fait qu'un signe n'est pas défini comme un sens déterminé pour pouvoir avoir son identité spécifique. » [Utaker, 1996 : 53]

Ce constat condamne-t-il toute analyse de la polysémie dans la perspective saussurienne ? Nous percevons deux directions qui permettent de conserver cette conception du signe et de poursuivre une analyse sémantique. La première interprétation de cette inter-dépendance stricte entre signifiant et signifié amènerait à considérer que le signifié serait un schéma abstrait, un « invariant » comportant des propriétés communes à tous les usages. C'est une piste explorée par les tenants de la psychomécanique [Guillaume, 1973], de l'énonciativisme [Culioli, 1990], et de la sémantique cognitive [Langacker, 1987]. L'autre perspective est de considérer que le contexte agit comme l'environnement structurant dans lequel se meut le signe et que l'interprétation du signe dépend du type de contexte dans lequel il est employé : la variation de sens s'explique par la variation du contexte.

L'absence de polysémie peut tout d'abord être justifiée par l'absence de sentiment de polysémie. On peut reprendre l'observation suivante de M. Bréal, qui a forgé cette notion :

> « Quand nous voyons le médecin au lit d'un malade, ou quand nous entrons dans une pharmacie, le mot *ordonnance* prend pour nous une couleur qui fait que nous ne

> pensons en aucune façon au pouvoir législatif des rois de France. Si nous voyons le mot *Ascension* imprimé à la porte d'un édifice religieux, il ne nous vient pas le moindre souvenir des aérostats, des courses en montagne, ou de l'élévation des étoiles. On n'a même pas la peine de supprimer les autres sens du mot : ces sens n'existent pas pour nous, ils ne franchissent pas le seuil de notre conscience. Il en doit être ainsi, l'association des idées se faisant heureusement chez la plupart des hommes d'après le fond des choses, et non d'après le son. » [Bréal, 1897 : 157]

Nous avons également vu que certains phénomènes polysémiques pouvaient s'expliquer, au moyen du principe de métonymie intégrée (cf. *infra* 1.2.5), par des proximités référentielles entre objets (relation partie tout). Plus précisément, ce principe revient à utiliser des propriétés référentielles qui expliquent qu'un terme soit employé dans un « sens » autre que l'objet qu'il dénote, sans pour autant multiplier les significations de cette dénomination. En résumé, ce principe tend à réduire l'existence de la polysémie (ou à la redéfinir comme une métonymie). L'inexistence de la polysémie est une position défendue par F. Rastier. Il affirme que la polysémie est un artefact de la linguistique. Rastier nous rappelle à ce sujet une des distinctions fondamentales des approches linguistiques, nommément sémasiologique et onomasiologique :

> « On sait que la sémasiologie prend le signifiant pour invariant, et considère le problème de la polysémie comme fondamental, alors que l'onomasiologie part du signifié, et considère la synonymie comme primordiale. » [Rastier, 2003 : 41, note 6]

D'après Rastier, la polysémie procède d'une décontextualisation indue des mots à l'étude :

> « pour constituer ou constater la polysémie, on ne tient pas compte des contextes, puisqu'on juxtapose des acceptions qui n'ont pas la même histoire, ne se trouvent ni dans les mêmes discours, ni dans les mêmes genres, ni souvent dans les mêmes textes. » [Rastier, 2004 : 31]

Cette position contextualiste interprète les hypothétiques unités polysèmiques comme des signes différents, une fois leur origine (con)textuelle caractérisée. Elle sous-tend un postulat fort de départ, qui est le suivant :

v. le contexte détermine totalement le sens d'un mot (« le global détermine le local »)

Il serait néanmoins injuste d'imputer aux défenseurs de la polysémie une cécité au contexte, car la majorité des études cherchent justement à faire correspondre des sens à des contextes précis : soit par exemple par les structures syntaxiques dans lesquelles est employé un mot, soit encore par les contraintes sémantiques de l'environnement propositionnel (structures prédicatives par exemple). Le Lexique Génératif [Pustejovsky, 1998] est un exemple de modèle de composition sémantique permettant de traiter les cas de polysémie systématique (qui ne dépend pas d'un genre ou discours).

Si le contexte détermine le sens d'un mot, il faut en toute logique caractériser ce contexte. Comme celui-ci est en premier lieu constitué d'unités linguistiques (le texte), il faut analyser les relations sémantiques entre contexte(s) et mot. Il est donc nécessaire de disposer au préalable de connaissances sur le sens des mots, sans quoi on ne peut analyser le sens, ni répondre au problème de la polysémie (qu'on nie son existence ou pas). Mais le contexte n'est pas uniquement lexical comme nous allons le voir, en présentant les critères de typologie textuelle proposés par Rastier.

1.3.3. Définir le texte

Pour aborder la linguistique du texte, Rastier (voir [Halliday & Hasan, 1976], [Halliday, 1994] pour une autre conception de la linguistique textuelle) propose tout d'abord de distinguer les différents niveaux d'analyse sémantique d'un texte :

- le niveau du mot ou plus exactement de la lexie (microsémantique),
- le niveau de la phrase ou de l'énoncé (méso-sémantique)
- et le niveau du texte (macrosémantique).

Tous ces niveaux (qui peuvent chacun être décomposées en sous-niveaux) sont liés et interdépendants.

La théorie de Rastier est sémique, c'est-à-dire qu'elle décrit les textes en fonction de sèmes. Elle est également différentielle car ces sèmes sont obtenus par comparaison des signes (mise à jour de leurs différences). Le sème est la plus petite unité de sens (qu'on peut concevoir comme une propriété, un trait). Par exemple, /haut/ dans *plafond* par opposition à /bas/ dans *parquet*. Les sèmes sont combinés dans des sémèmes qui correspondent au signifié d'une unité lexicale. Il est ainsi possible d'établir que certains sèmes sont spécifiques à un sémème vis-à-vis d'une classe d'unités lexicales. Par exemple, en comparant *homme* et *femme*, on peut isoler le sème spécifique /masculin/ pour le sémème de *homme*. Parallèlement on peut identifier le sème générique dans le cadre de cette comparaison, qui est commun aux sémèmes de ces deux unités lexicales, le sème générique /personne/ ou /humain/. Comparer des unités lexicales (ou plutôt leurs sémèmes) peut se faire selon trois classes ou niveaux de généralité sémantique : on peut identifier des sèmes génériques,

- du point de vue du taxème, la classe minimale de sémèmes (par exemple *cuillère* et *fourchette* forment un taxème où le sème générique est /couvert/),
- du point de vue du domaine, classe intermédiaire qui se caractérise par rapport à une pratique sociale (/alimentation/ pour *fourchette*, /maritime/ pour *bateau*, etc.)
- et du point de vue de la dimension, qui concerne les distinctions les plus génériques (/humain/ pour *cuisinier*, /concret/ pour *cuillère*).

Enfin, les sèmes peuvent être inhérents (hérités par défaut de la langue, ou « définitoires » ; /noir/ pour *corbeau*), afférents contextuels (propagés dans le contexte, dû à des associations répétées) ou encore afférents socialement normés (connotés socialement ; /péjoratif/ pour *corbeau*).

Ces distinctions, comme nous venons de le dire, concernent les (signifiés des) unités lexicales, mais le fait que leurs sèmes n'émergent que par comparaison, implique leurs co-occurrence dans un espace cohérent qu'est la phrase (mésogénérique) et/ou le texte (macrogénérique). Le sens ne peut être analysé qu'au sein du texte (ou d'un ensemble de textes proches). Par exemple, on n'aurait pas idée de comparer *bateau* à *fourchette* si leur co-présence dans un texte n'était pas attestée (voire signifiante …). On comprend que l'analyse sémique dépend crucialement de la prise en compte de la nature du texte (ou du type si l'on préfère) : les sèmes qui émergeront d'un texte de type culinaire, seront différents de ceux qui émergeront d'un texte de type immobilier. Ainsi le sémème de *fourchette* ne sera pas caractérisé par les mêmes sèmes en (17) et en (18)

(17)Piquer les saucisses avec une fourchette.
(18)Plus de la moitié des projets immobiliers se situent dans une fourchette de 100.000 à 299.000 €.

Il ne fait aucun doute que ce sont deux signes différents dans le cadre de la sémantique textuelle. On peut, certes, chercher des sèmes communs à ces deux usages de la même forme, un invariant dont la définition pourrait probablement être fournie, ou expliquer les différentes transitions de sens qui aient pu mener de l'un à l'autre, mais nous manquerions de voir que chacun de ces usages désigne des référents bien différents. En (17), fourchette est un ustensile de cuisine qui pique, alors qu'en (18), fourchette est un intervalle quantificationnel qui s'applique à deux valeurs extrêmes, monétaires par exemple, ou à des valeurs conçues d'un point de vue numérique (que dire de la fourchette aux jeux d'échecs ? Au rugby ? En anatomie ?).

Mais Rastier ne se limite pas à cette seule analyse, puisque, étant donné que les sèmes varient en fonction des textes, il faut également pourvoir à une théorie de la structuration des textes. On a coutume en linguistique textuelle de distinguer des genres, comme l'article scientifique, le roman, la conversation téléphonique, la lettre professionnelle, et ainsi de suite (par exemple [Adam, 2001]). Il est possible de concevoir de nombreuses catégories de genres ainsi que de sous genres. Plutôt que de donner une typologie de ces genres, Rastier préfère proposer des critères sémantiques pour caractériser les textes. Il distingue quatre composantes et la première est la plus étudiée :

1. *La Thématique* : elle permet de caractériser le thème (dit « isotopie »), ou contenu d'un texte en s'appuyant sur la récurrence de sèmes des unités lexicales du texte (ou « sème isotopant »). Les isotopies sont génériques ou spécifiques, selon la distinction générique/spécifique élaborée pour les sèmes, et en fonction des trois classes : taxème, domaine, dimension. Une combinaison récurrente de sèmes s'appelle une « molécule sémique ». La Thématique permet *in fine* de rendre compte de la cohérence sémantique des textes.
2. *La Tactique* : elle s'intéresse aux structures linéaires qui composent un texte du point de vue du sème. Elle concerne les notions de position (initiale, finale, etc.), de disposition et de progression linéaire qui peuvent être le produit d'une figure de style (parallélisme, chiasme, etc.) ou de structures plus normatives (ordre de présentation d'un récit par exemple). Le choix du terme *tactique* pour désigner cette composante renvoie à l'identification d'une stratégie de disposition des unités sémantiques.
3. *La Dialectique* : Elle étudie, comme la Tactique, les unités sur un plan séquentiel, mais du point de vue de leur représentation, c'est-à-dire en tant qu'ils constituent des événements, des épisodes, etc. Elle se divise en deux niveaux, événementiel et agonistique.
    - Le niveau événementiel définit trois unités : acteur, fonction et rôle. L'acteur est une entité qui accumule les sèmes, génériques et spécifiques, par lesquels il est régulièrement désigné (dans toutes ses manifestations et désignations du texte). Les fonctions désignent des interactions typiques entre acteurs, soit l'ensemble des processus (événements et états) auxquels ils peuvent être associés. Elles se caractérisent également par l'accumulation de sèmes génériques et spécifiques de ces processus, indépendamment de l'identité de ses acteurs. Les fonctions peuvent être regroupées en syntagmes fonctionnels ou scripts (scénarios). Les rôles font référence aux rôles thématiques récurrents (la théorie des cas ; [Tesnière, 1965], [Fillmore, 2003]) qu'exercent les acteurs dans le cadre des processus.
    - Le niveau agonistique est plus abstrait car il se conçoit comme la projection du niveau événementiel sur le plan inter-textuel (ou au sein d'un texte suffisamment complexe) : il comporte deux unités qui caractérisent les similarités entre fonctions et entre acteurs et qui peuvent se manifester dans plusieurs textes. Un agoniste se définit en fonction des

sèmes spécifiques et/ou des rôles, communs à un groupe (classe) d'acteurs. Cette unité permet de rendre compte de similarités entre acteurs sans avoir à s'appuyer sur leurs sèmes génériques. Une séquence résulte de l'analyse de la récurrence de configurations de fonctions dans un texte. Elle permet de décomposer un texte en fonction de blocs ou épisodes qui ont un même « objectif » (argumentatif, descriptif, etc.) et d'analyser la manière dont ces séquences sont articulées.

4. *La Dialogique* : Elle concerne l'étude des acteurs d'un texte et permet de structurer la complexité des énoncés qui leur sont attribués. Il ne s'agit plus ici de caractériser les événements auxquels un acteur peut être associé, mais ses jugements, croyances et connaissances que Rastier regroupe sous le nom d'« univers ». Pour chaque acteur, cet univers peut être organisé en trois types de « mondes ». Le monde factuel concerne tout ce qu'il énonce, ou affirme, considère « vrai » (à propos d'autres acteurs ou autre) ; le monde contre-factuel, ce qui est impossible ou irréel ; le monde possible, par tautologie, ce qui est possible. Si certaines classes d'acteurs peuvent être énonciateurs, tous ne le sont pas. Enfin, ce qui est considéré comme vrai pour tous les énonciateurs fait partie de l'univers de référence du texte, qui est attribué par défaut à « l'énonciateur représenté » (le narrateur omniscient par exemple).

La sémantique textuelle ainsi esquissée offre, de toute évidence, un cadre d'analyse riche et justifie, de notre point de vue, la prise en compte, ne serait-ce que partielle, des contraintes ou structures textuelles. Néanmoins, si toute analyse sémantique ne peut ignorer le cadre textuel dans lequel elle progresse, notre objectif n'est pas de caractériser les textes, mais d'analyser les mots d'un point de vue lexical. De fait, nous devons tenter d'identifier plus précisément les incidences du texte et de ses propriétés sur notre entreprise. Elle permettra, nous l'espérons, de mieux mettre en lumière les rapports entre sémantique référentielle et sémantique textuelle et d'initier une sémantique référentielle consciente des enjeux textuels.

1.4. Epilogue : Intégration réinterprétation et synthèse

1.4.1. Sémantique textuelle et paradigme référentiel

Le rapprochement que nous souhaitons établir prend pour point de départ le constat que fait Kleiber à propos de la nature « existentielle » du sème :

> « Par ailleurs, et c'est le reproche de fond qu'on peut émettre contre toute théorie sémantique uniquement différentielle, elle n'est en elle-même pas apte à dire quel est le sens d'une unité. En apparence, les sèmes sont dégagés par l'opposition des lexèmes entre eux, mais, en réalité, la connaissance de la signification de chacun de ces lexèmes doit précéder leur confrontation. Dans le cas contraire, on ne peut mettre au jour aucun trait de signification pertinent. » [Kleiber, 1999a : 38]

Cet argument rappelle notre remarque (en 1.1) qu'un référentiel interprétatif (dans ce cas le réel) est nécessaire dans toute sémantique. Kleiber veut ici démontrer la primauté du réel (tel qu'il le redéfinit) sur toute interprétation. Qui plus est, pour comparer deux sémèmes, il faut savoir de quoi l'on parle : deux signes partagent des sèmes génériques uniquement parce qu'ils désignent chacun un référent (ou une classe de référents). Par exemple, les sémèmes de *fourchette* et de *cuillère* ne seront pas inclus dans un même taxème si *fourchette*, dans le texte, désigne un intervalle numérique. Ces rapports doivent être « perçus comme significatifs ».

Il est également crucial de pouvoir caractériser les types de relation que peuvent entretenir deux ou plus signifiés : dire que l'on peut comparer tout ce qui co-occurre dans un texte est insuffisant. Tel et tel rapport établi entre deux sémèmes qui fait émerger tel ou tel sème, est significatif relativement à un type de rapport. Par exemple, les sémèmes de *fourchette*, *couteau* et *cuillère* semblent entretenir le même type de rapport que ceux de *train*, *voiture*, *métro*, *avion* et *bateau* : ils font partie d'une catégorie (les ustensiles et les véhicules). Ce type de rapport associatif n'a rien à voir avec les rapports syntagmatiques entretenus par les sémèmes de *je*, *vais*, *à* et *Paris*. On se rend également compte que le sème et le rapport changent en fonction du nombre de signes considérés (*train* et *métro* peuvent partager le sème /*véhicule-sur-rail*/, mais pas *train*, *voiture*, *métro*, *avion* et *bateau*). Pourrait-on comparer tous les mots d'un texte ? Non, il semble qu'on ne puisse pas comparer *bateau* et *fourchette* uniquement parce qu'ils co-occurrent dans un même texte, s'il n'y a pas de rapport d'ordre paradigmatique ou syntagmatique et si ce rapport n'instancie pas un type de relation. En (19), par exemple la comparaison entre *fourchette* et *bateau* ne nous apprend véritablement rien sur leur signification, mais plutôt sur leurs propriétés dialectiques (ils participent à un même événement).

(19)J'ai dû perdre ma fourchette sur le bateau.

Ce n'est que parce que l'on suppose que *fourchette* est un type d'objet matériel et que *bateau* est un véhicule sur lequel on peut se déplacer, que l'on peut tenter d'interpréter cette phrase, comme dans le contexte d'une déclaration de la perte d'un objet lors d'un passage effectué à bord d'un bateau.

Kleiber nous invite également à nous interroger sur la substance sémantique de ce mot entre barres obliques (/sème/) : qu'évoque-t-il ? Quel sens du mot entre barres obliques doit-on prendre comme sème ? Ses propres sèmes ? Génériques, spécifiques ? Par rapport à quel taxème ?

Du point de vue du paradigme référentiel, s'il est vrai qu'au fond, Kleiber redéfinit la notion d'existence à, non pas ce qui est concret, mais tout ce à quoi on peut référer, c'est-à-dire en intégrant

l'imaginaire et l'abstrait, ou tout ce qui est « non-concret », des questions subsistent quant aux incidences de cette appropriation. La démonstration de Kleiber consiste principalement à justifier le fait que, peu importe ce que l'on analyse, il est toujours possible (nécessaire ?) de retomber sur ses « pieds référentiels ». Mais comment analyser d'un point de vue référentiel, des référents abstraits [Kleiber, 1994] ou fictifs ? Sur quels référentiels interprétatifs doit-on se situer ? Doit-on lier systématiquement *in fine* notre interprétation à la réalité, ou est-il possible d'accorder une forme d'autonomie et de cohérence interne à d'autres mondes interprétatifs ? Si oui, quelle est leur nature et comment les décrire ? Le texte peut alors proposer des pistes.

De son côté, la sémantique textuelle ne remet en cause ni l'existence ni l'importance de la référence, son théoricien s'attaque plutôt à ce qu'il considère comme une vision simpliste du langage, qu'il résume par le terme *tradition «* logico-grammaticale », qui est locale, c'est-à-dire qu'elle ne prend pas suffisamment en compte la « globalité » dans laquelle les énoncés sont employés (produits et interprétés). Mais cette tradition logico-grammaticale est bien trop caricaturale comme en attestent les travaux en sémantique référentielle sur le contexte et sur le texte. Rien n'empêche, *a priori*, de concevoir une sémantique textuelle de la référence, ni non plus une sémantique référentielle du texte. En tant que théorie herméneutique, la sémantique textuelle se doit d'intégrer la référence ainsi que de la distinguer du langage. Voici comment Rastier conçoit son interaction :

> « Ce que nous appelons ici *référence* n'est pas un rapport de représentation à des choses, mais un rapport entre le texte et la part non linguistique de la pratique où il est produit et interprété.[...] La référence ainsi définie ne relève pas de la représentation mais de l'action, telle qu'elle est structurée par une pratique. Plus généralement, l'ordre référentiel met en jeu, de façon différenciée, au sein de chaque pratique sociale, les rapports variables entre la sphère sémiotique (ici les suites linguistiques), la sphère des représentations (ici les impressions référentielles), et la sphère physique (ici les « objets »). Il faut donc préciser les modes de référenciation propres aux pratiques sociales. » [Rastier, 1996 : 34]

Rastier critique la vision simpliste qui consiste à associer signe à référent pour au moins deux raisons. En premier lieu, Rastier nous rappelle que ce sont les syntagmes, c'est-à-dire le produit d'une construction linguistique, et non les signes qui peuvent être dits référents :

> « En somme, aucun signe linguistique ne "réfère", parce que la propriété de susciter des images mentales est propre aux syntagmes (dont le mot) mais non à chacun des signes qui les [constitue]. Il faut donc que des conditions contextuelles soient remplies pour qu'un morphème puisse participer à l'élaboration d'une image mentale. » [Rastier, 1989 : 253]

En second lieu, analyser la référence suppose également d'analyser les relations entre les référents, *comme si* la référence ne pouvait véritablement prendre sens, ou n'être expliquée, qu'au sein d'un texte et à travers ses composantes :

> « La sémantique textuelle devra, au-delà du mot et du syntagme, traiter de la composition des impressions référentielles : elle dépend des quatre composantes sémantiques. Toutefois, la cohésion des impressions référentielles ne détermine qu'un des aspects de la textualité. » [*ibid.*]

Nous nous intéressons à cet aspect de la textualité : comment une perspective globale peut contribuer à une meilleure appréhension des référents ? Il nous semble que la différence fondamentale entre ces deux traditions réside sur le fait que l'une focalise sur l'unité (le référent, sa classification, ses propriétés, etc.), alors que l'autre donne plus de poids à la structure (les isotopies, les séquences, les scripts, les mondes, les progressions, les stratégies, etc.). Notre question peut donc se reformuler ainsi : quelles structures textuelles peuvent être associées au référent ? Ces structures peuvent-elles être (re)définies référentiellement ? Nous proposerons dans la partie suivante des pistes pour l'intégration et la réinterprétation de ces composantes dans le cadre d'une sémantique référentielle.

1.4.2. Macrosémantique du référent : propositions

Pour établir un parallèle entre les deux traditions que nous avons présentées, nous devons tout d'abord identifier le point de rencontre à partir duquel on pourra transférer les concepts de la sémantique textuelle à la sémantique référentielle. Il se place à notre avis entre le sème et l'attribut. Les sèmes, comme les attributs, sont les unités définitoires d'une dénomination. Nous proposons donc d'associer sème à attribut :

vi. Un sème correspond à l'attribut d'une dénomination.

Les sèmes sont des propriétés des référents, et non plus uniquement des sémèmes. Les dénominations (nom propre ou pas) peuvent dénoter et connoter selon l'usage qui en est fait en contexte et selon les connaissances qu'on en a. Dans ce contexte, nous nommerons ces propriétés ou sèmes des catégories sémantico-référentielles. Ces catégories correspondent à l'ensemble des connaissances (de tout type) que l'on peut acquérir sur un objet et qui caractérisent le sens de l'unité linguistique qui dénote cet objet.

Nous devons tout d'abord rappeler que les connaissances sont toujours situées.

- Comme nous l'avons vu à travers l'analyse de la logique de Mill, la connotation d'une dénomination est tributaire d'un objectif de classification. Parallèlement, les unités linguistiques sont toujours ancrées dans un contexte et il importe d'analyser les liens qu'elles tissent avec leur environnement textuel. Plus précisément, la caractérisation d'une dénomination repose sur le contexte de production (le contexte dans lequel elle est employée) et sur le contexte de réception (le point de vue adopté). Le contexte de production peut être appréhendé par les catégories socio-discursives auxquelles appartient un texte (genre, style, domaine, thème, etc.) et le contexte de réception peut l'être par la grille de lecture de ce texte ainsi que par l'objectif de son usage.
- Le contexte de production d'un texte influence l'interprétation : les mots qui y sont employés ont un sens particulier, renvoient prioritairement à des référents de ce contexte. Dans un domaine comme la médecine, sur un sujet comme les relations médecin-patient, la connotation d'une dénomination comme *ordonnance* renverra à des référents particuliers. En d'autres termes, la référence dépend du contexte de production.
- Mais le contexte de réception influence également l'interprétation. Pour interpréter « correctement » un texte ou une dénomination, il faut posséder les connaissances requises sur leur usage. Dans le cas contraire, on risque de mal identifier le référent. Un tel décalage peut mener à d'autres interprétations ou à des contre-sens. Par conséquent, le contexte de production n'existe qu'à travers le prisme d'un contexte d'interprétation ; ce dernier l'articule, le contexte de production n'existe pas en soi ou indépendamment.

1.4.Epilogue : Intégration réinterprétation et synthèse

Dans notre exposé des composantes macro-sémantiques (cf. *infra* 1.3.3), nous avons cherché à extraire pour chacune leur distinctivité. Or, on se rend compte que dans leur application, les distinctions ne sont pas scrupuleusement respectées : Rastier argumente en faveur d'une interaction des composantes, en fonction de niveaux et de types de normes [*ibid.* : 103–109], tout en leur attribuant une unité, une indépendance. Pourtant, l'auteur propose dans un article sur la thématique de l'ennui [Rastier, 1995 : 224–228] une description de phénomènes qui auraient plus leur place dans la dialectique, telle que nous l'avons présentée. Ailleurs, il parle de la « structure thématique d'un acteur » [Rastier, 1989 : 274–275, note 40]. En tant qu'observateur critique plutôt qu'expert praticien de la sémantique textuelle, cette articulation entre thématique et dialectique nous paraît problématique (on peut l'imputer à l'extrême complexité qu'une théorie du texte doit posséder).

En effet, il nous paraît nécessaire de distinguer *thème* et *acteur* aussi clairement que possible, pour mieux apprécier leurs contributions, surtout que leur substance sémantique s'établit, pour tous deux, sur la combinaison (récurrente) de sèmes.

- La catégorie d'acteur est judicieuse car elle permet de regrouper toutes les occurrences d'un même référent dans un texte ; elle est utile à une articulation entre texte (les occurrences) et référence (le type). Du point de vue de la sémantique référentielle, on ne considère pas la diversité des occurrences d'un référent dans un texte comme une dimensions pertinente à sa définition (mais voir [Kleiber, 1994 : 147]) : il est les deux à la fois.
- Nous pensons qu'il devrait en être de même pour les fonctions. Rastier pose une distinction de nature entre acteur et fonction (par « homologation » du niveau méso-sémantique), mais, alors qu'il considère que l'acteur peut cumuler des sèmes, il considère les fonctions comme des classes de processus et par conséquent, ne se donne pas la possibilité de considérer formellement un processus comme un référent pouvant se répéter dans le texte. C'est pour saisir cette possibilité que nous aurons recours à la catégorie de fonction. Nous devrons donc réorganiser ce niveau dialectique.

Nous proposons donc de restreindre l'acteur par un principe d'identité référentielle, et, à l'inverse, d'attribuer au thème la caractérisation des associations qui peuvent être identifiées malgré une hétérogénéité référentielle ; ce que nous traduisons par les principes suivants :

vii. Les acteurs et les fonctions sont des catégories regroupant la totalité des occurrences d'un même référent (entité ou processus)

viii. Les thèmes sont des catégories regroupant les récurrences sémantiques s'établissant entre les attributs de référents différents

L'univers de référence majeur des acteurs, fonctions et thèmes est le texte, mais il peut s'étendre au-delà : plusieurs textes différents peuvent traiter du même référent. Le texte est le milieu d'étude privilégié des référents.

Pour mieux caractériser les référents, nous souhaitons également marquer une distinction de niveaux comme le fait Rastier pour la composante dialectique. Nous reformulerons son modèle en trois niveaux :

(1) Le niveau d'identité référentielle : les acteurs et les fonctions correspondent aux réseaux d'occurrences du même référent dans un ou plusieurs textes.

(2) le niveau d'interaction référentielle : il concerne les interactions (d'ordre syntagmatique) entre référents, dans une même proposition (les entités et processus), ou dans un même

texte (les acteurs et fonctions).

(3) le niveau de classification référentielle : il correspond aux relations sémantiques (d'ordre paradigmatique) entre éléments d'une même catégorie, processus, référents, acteurs ou fonction.

1. Identité référentielle

L'acteur « n'est » pas forcément une personne, et n'est pas forcément animé [Rastier, 1989 : 73] : il participe à un événement ou à une situation dans laquelle c'est un référent. Les sèmes qu'accumulent cet acteur peuvent être également (et le sont certainement) pris dans des structures thématiques (au sens d'isotopie), mais ces sèmes sont reconsidérés du point de vue d'une id-entité qui est référentiellement stable (et nous éviterions dans ce cas de parler de structure thématique d'un acteur). Cette identité n'est pas réservée aux individus : les classes d'acteurs seront également des acteurs, bien qu'on ne puisse pas avec précision identifier les référents. Les événements peuvent également être des acteurs : les expressions comme *le 11 Septembre*, *le procès D'Outreau*, instancient des catégories d'acteur de type d'événement. Mais les processus (les verbes finis) sont beaucoup plus fréquents dans cette catégorie. On peut résumer ce niveau par la figure (1.5).

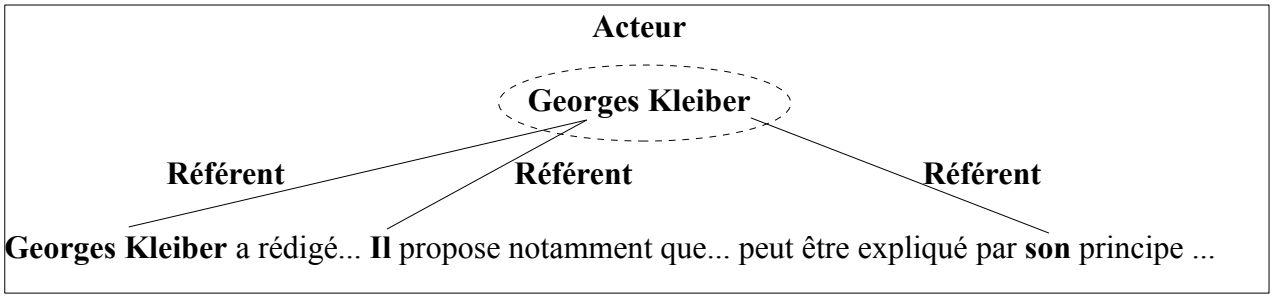


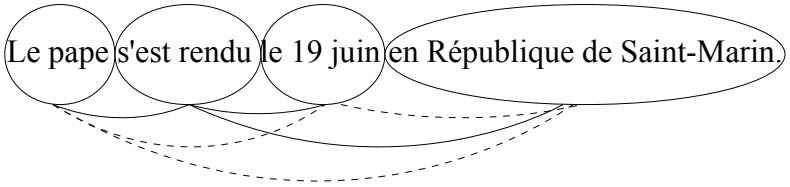
FIG 1.5 – Occurrences et Identité du référent

On peut concevoir l'acteur comme un index (dénotatif) de l'ensemble des occurrences d'un référent. Les processus sont des référents plus éphémères, en partie du fait qu'ils dépendent d'autres entités pour exister (mais on peut considérer *il pleut* comme des processus sans entité), ils peuvent être caractérisés par rapport à la durée et à la date. Les fonctions désignent des récurrences de processus, pas nécessairement associés aux mêmes acteurs. Les phrases habituelles (*Paul va normalement à l'école à pied*) illustre un processus qui est aussi fonction (voir à ce propos [Kleiber, 1987]), mais on peut composer une fonction en associant les occurrences d'un même processus.

2. Interaction référentielle

Un acteur est toujours situé dans un contexte où il interagit avec d'autres référents. L'acteur remplit des rôles et participe à des processus qui le localisent notamment du point de vue spatio-temporel. Il est également caractérisé par rapport à d'autres acteurs, que les processus réunissent. Cette propriété traduit la notion de *dépendance référentielle* [Garde, 1985]. Dans l'exemple (20), on pourra identifier une dépendance référentielle entre chacun des référents.

(20)



La récurrence de ces dépendances référentielles pourront être étudiées dans un ou plusieurs textes. Il est aussi utile de pouvoir caractériser le rôle d'un référent vis-à-vis d'un autre. Les rôles sémantiques ont été largement décrits en linguistique (agent, patient, localisation spatiale, localisation temporelle, etc.), ils correspondent aux rôles d'un référent vis-à-vis d'un processus. Les rapports entre les acteurs d'un même processus (en pointillés sur l'exemple 20) sont moins souvent décrits.

3. *Classification référentielle*

Les référents ne sont pas uniquement situés dans une interaction, ils sont enfin qualifiés et catégorisés. La catégorisation, telle que nous l'avons présentée dans le cadre du paradigme référentiel, peut prendre différentes formes dans un texte, mais c'est le texte qui nous livre les clés de la catégorisation. Les expressions linguistiques apportent tout d'abord des informations sur le référent, comme le nom et la fonction professionnelle en (21).

(21)*Le professeur de linguistique André Martinet* est mort vendredi 16 juillet 1999

On peut également trouver ces informations en apposition (*André Martinet, professeur de linguistique, ...*) ou encore en relation attributive (*André Martinet est professeur …*). Les acteurs peuvent être qualifiés (*Paul est heureux*). On peut enfin proposer les relations partie - tout (méronymie), comme des relations de catégorisation. Rappelons qu'elles s'appliquent autant aux acteurs-individus qu'aux acteurs-classe.

Ces informations contribuent à nos connaissances des acteurs et nous permettent également de comparer des acteurs en fonction de ces catégories. Le système des catégories est complexe et varie en fonction du point de vue et du texte à l'étude (contextes de réception et de production).

Pour ce qui concerne les processus et les fonctions, on pourra s'inspirer des cadres sémantiques de Fillmore et notamment des travaux réalisés sur FrameNet. Les cadres sont des « familles » de processus. Ils ne sont en effet, pas définis par rapport à l'identité référentielle mais respectivement à une notion, comme le déplacement par exemple : ce sont des catégories qui peuvent s'organiser de manière hiérarchique ou « associative », certains sont des parties de cadres plus complexes comme les scénarios (les séquences de Rastier).

Nous pourrions proposer d'autres éléments pour intégrer référence et texte, en nous interrogeant sur l'intérêt des trois autres composantes textuelles, mais nous nous contenterons de cette courte présentation. Dans ce chapitre, nous avons tenté d'illustrer la complexité et les soubassements d'une étude du sens dans une perspective linguistique. Nous avons mis en valeur l'importance d'une réflexion sur le rapport entre langage et réalité ainsi qu'entre langage et texte. Le langage est l'instrument de catégorisation par excellence et les textes constituent des réservoirs de connaissance sur la réalité. Comme nous le verrons, la projection de référents sur des expressions linguistiques est mise à mal par les subtiles fluctuations du sens en contexte. Pour leur attribuer une texture référentielle, les mots sont mieux appréhendés par l'usage qui en est fait dans le texte. Cette

étude de l'usage des mots peut être doublement enrichissante : d'une part, elle peut permettre de saisir avec une meilleure précision la contribution sémantique de chaque mot, et d'autre part, elle peut faire émerger de nouvelles connaissances sur ces mots et donc leurs référents. L'approche que nous développerons sera guidée par des méthodes quantitatives inspirées de la linguistique de corpus : nous nous intéresserons aux catégories et relations sémantiques observées à l'échelle d'un corpus.

2. Sémantique et linguistique de corpus

Un texte peut recevoir plusieurs définitions et désigner des objets très différents. Rastier le définit très largement comme « une suite linguistique empirique attestée, produite dans une pratique sociale déterminée et fixée sur un support quelconque » [Rastier, 2001 : 21]. Une telle approche élargit le champ d'application de la linguistique à une multitude d'objets sémiotiques dont la taille, le support, le canal (oral/écrit) sont autant d'éléments de variation.

Ce chapitre présentera des méthodes d'analyse sémantique du texte, en ne se consacrant qu'à un courant, celui de la linguistique de corpus britannique (le terme de « linguistique de corpus informatisée », LCI, serait plus juste). Ce courant a développé des pratiques d'analyse empirique du texte assistée par ordinateur qui ont principalement été appliquées à la création de dictionnaires et de grammaires de langue. En France, J. Léon [Léon, 2008] a contribuer à mieux saisir les spécificités de cette « école contextualiste », en étudiant particulièrement les travaux de J. R. Firth (voir également [Williams, 2006] pour une description historique et [Legallois, 2004] pour une application en TAL). Nous espérons à travers ce chapitre, tout d'abord, souligner les points de controverse, mais surtout, relever les points névralgiques de ce courant en analysant les travaux de J. McH. Sinclair.

2.1. La Linguistique de corpus

2.1.1. Corpus et intuition

La linguistique de corpus n'aurait pas vu le jour sans l'informatique. Mais l'informatique a également permis le développement d'autres domaines de recherche linguistique, dont le Traitement Automatique des Langues (TAL). Le TAL (qui englobe les deux approches « Linguistique Computationnelle » et « Ingénierie Linguistique » ; [Bourigault, 2007 : 25–42]) est un domaine essentiellement orienté vers la création d'applications informatiques qui s'appuient sur des connaissances linguistiques. Y est notamment évalué l'apport d'informations linguistiques dans l'accès à l'information. Parmi les applications concernées, le TAL contribue à la Recherche d'Information, l'Extraction d'Information, le Résumé Automatique, les Systèmes de Question-Réponse, la Traduction Automatique.

G. Leech nous informe que la LCI comme le TAL ne sont pas des disciplines au sens strict du terme, mais se définissent par les méthodes et outils qu'il mettent en place :

> « The only other branch of linguistics which, like corpus linguistics, refers to a tool or methodology rather than a subject-matter is computational linguistics, defined as the investigation of language by means of computers. But nowadays, there is an obvious and growing overlap between corpus linguistics and computational linguistics. When we talk about corpus linguistics today, of course we assume that the corpus is machine-readable, and is to be investigated by means of computers. So in fact the branch of linguistics we are discussing at this Symposium should strictly be labeled “computer corpus linguistics” to distinguish it from the corpus linguistics of the pre-computer age. » [Leech, 1992 : 106]

La LCI et le TAL partagent des objets similaires et emploient des outils communs : corpus, programmes informatiques ou ressources linguistiques. N. W. Francis (voir aussi [Laks, 2008]) remarque que l'usage du corpus en linguistique n'est pourtant pas né de l'invention de l'ordinateur :

> « I will confine myself to corpora accumulated B.C., i.e. before the use of computers [...] Some seem to believe that there were not corpora before that. The truth is that many important corpora of English were assembled long before the computer was invented. » [Nelson W. Francis, 1992]

Si les travaux sur corpus existaient déjà, en quoi l'informatisation des corpus justifient la distinction d'une nouvelle linguistique de corpus ?

Tout d'abord, considérer que l'informatisation de l'objet d'étude du linguiste permet uniquement de faciliter son travail serait réducteur, comme l'affirme Leech :

> « In my contribution to the Symposium, I wish to argue that computer corpus linguistics (henceforth CCL) defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. The computer, as a uniquely powerful technological tool, has made this new kind of linguistics possible. So technology here (as for centuries in natural science) has taken a more important role than that of supporting and facilitating research: I see it as the essential means to a new kind of knowledge, and as an “open sesame” to a new way of thinking about language. » [Leech, 1992 : 106]

2.1.La Linguistique de corpus

L'existence de corpus informatisés permet d'envisager de nouvelles approches du langage et permet de définir de nouvelles hypothèses de recherche. Un apport majeur de cette synergie entre linguistique et informatique est la possibilité de quantification de phénomènes linguistiques au sein du cadre qu'est le corpus. Le type d'entreprise scientifique qu'envisage Leech dans cet article est la constitution de modèles statistiques de la performance linguistique (en tant que produit et non en tant que processus ; [*ibid.* : 108]). J. McH. Sinclair voit dans l'existence de corpus informatisés, une opportunité unique de développer de nouvelles techniques de description du langage :

> « The most exciting aspect of long-text data-processing, however, is not the mirroring of intuitive categories of description. It is the possibility of new approaches, new kinds of evidence, and new kinds of description. Here, the objectivity and surface validity of computer techniques become an asset rather than a liability. Without relinquishing our intuitions, of course, we try to find explanations that fit the evidence, rather than adjusting the evidence to fit a pre-set explanation. » [Sinclair, 1991 : 36]

Les outils informatiques sont des atouts (« asset ») et les données attestées sont des preuves (« evidence ») qu'il faut expliquer.

Face à cet enthousiasme partagé pour découvrir de nouvelles « frontières » de la recherche, il nous faut émettre quelques réserves préalables sur l'impact de l'outillage informatique dans des disciplines qui n'y avaient pas alors accès, comme la linguistique, plus généralement les sciences du langage et de manière globale les sciences humaines. En effet, si, certes, on assiste à un bouleversement des méthodes d'accès aux données linguistiques, on est en droit de s'interroger sur l'avancement de problématiques propres à la discipline visée, sur le dépassement de difficultés ou sur la découverte de nouveaux faits. Somme toute, la technologie ne permet pas de réinventer la discipline, puisque concernant la linguistique, il s'agit toujours de décrire le langage. La technologie peut en revanche révolutionner les pratiques d'une discipline (comme l'affirme Leech dans la citation précédente) et ses méthodes d'approche de l'objet.

La technologie ne se substitue pas non plus à l'homme comme l'explique Sinclair dans la citation précédente. Bien qu'il les oppose, Sinclair considère l'introspection et le corpus informatisé, comme deux types valables de preuves [*ibid.* : 37] : il privilégie la seconde dans les premières phases de l'analyse et la première dans les phases ultérieures. Il est important de noter que dès le départ, la LCI ne se conçoit pas comme une approche objective, purement empirique, car malgré le fait que les linguistes puissent désormais s'appuyer sur de larges ensembles de données attestées et opérer des caractérisations quantifiées de leurs objets, ces données nécessitent d'être interprétées au moyen de leur subjectivité. M. Scott et C. Tribble introduisent leur ouvrage en développant un argument similaire :

> « The process operates in two stages. First, all the effort of a concordancer or a word-listing application goes into reducing a vast and complex object to a much simpler shape. That is, a set of 100 million words on a confusing wealth of topics in a variety of styles and produced by innumerable people for a lot of different reasons gets reduced to a mere list in alphabetical order. […] The advantage comes in the second stage where one examines the boiled-down extract, the list of words, the concordance. It is here that something not far different from the sometimes-scorned “intuition” comes in. This is imagination. Insight. Human beings are unable to see shapes, lists, displays, or sets without insight, without seeing in them “patterns”. It seems to be a characteristic of the homo sapiens mind that it is often unable to see things “as they are” but imposes on them a tendency, a trend, a pattern. » [Scott & Tribble, 2006 : 6]

Cette citation fait écho aux réflexions présentées chapitre 1 (la réalité est une réalité perçue par exemple) et introduit la notion de patron, qui est d'importance majeure en LCI et sur laquelle nous aurons l'occasion de revenir. Les auteurs sont aussi convaincus que les objectifs fondamentaux de l'étude du langage n'ont pas changé (qui sont d'après eux, d'ordre linguistique, textuel, psychologique et culturel ; [*ibid.* : 6–8]).

Sinclair est un peu plus subtile dans sa conception de l'interaction entre intuition et données. Il remarque que les données que renvoie un programme informatique ne correspondent pas toujours aux attentes de l'analyste. C'est dans ce conflit, c'est pour cette raison, pour expliquer ces nouveaux modes de représentation de l'objet, que naît l'intérêt pour une approche guidée par le corpus :

> « The discrepancies between the way a computer can identify things in a text and the expectations of a person who knows the language of the text is well worth investigation ».[Sinclair, 1991 : 36]

Ce à quoi fait allusion Sinclair concerne les difficultés que rencontre le linguiste dans cet environnement inconnu. En effet, le linguiste manipule des concepts et des catégories scrupuleusement caractérisées (le morphème par exemple), qui s'appuient sur le sens plutôt que sur la forme. En revanche, les logiciels qui leur ont été conçus pour analyser les textes (les concordanciers par exemple) imposent un mode d'interrogation plus rudimentaire dont l'unité principale est une succession de caractères alphanumériques. L'application de modèles sémantiques ne peut pas s'établir directement et c'est ce que Sinclair trouve intéressant.

2.1.2. Deux perspectives d'usage du corpus

L'usage de corpus informatisés par rapport à l'usage de corpus « version papier » peut être résumé par le passage d'une situation à une autre :

- Lorsqu'il utilisait des corpus non-numériques, le linguiste était limité par les oublis, la taille (devoir tout lire pour détecter un phénomène particulier), recopier et compter.
- À présent qu'il dispose de logiciels d'accès à ces mêmes corpus, d'autres problèmes se posent ; il doit notamment s'interroger sur les formes que peut prendre son phénomène avant de soumettre sa requête.

Ces problèmes liés à l'usage de l'ordinateur permettent de définir deux stratégies principales, peu importe le phénomène en jeu ou le niveau linguistique concerné :

- Améliorer les modes d'accès aux données pour permettre une interrogation linguistique de corpus

  La première stratégie consiste à créer des ressources complémentaires au corpus, contenant les informations connues du linguiste. L'enrichissement de corpus en est un exemple : les corpus sont annotés par des informations comme les catégories grammaticales et les lemmes (ou forme canonique d'indexation d'un dictionnaire : infinitif d'un verbe par exemple). Les interrogations du corpus deviennent plus complexes et s'appuient sur d'autres éléments que les formes de surface. La conception d'outils d'annotation automatique est une des contributions du TAL, bien qu'ils ne soient pas toujours taillés aux besoins du linguiste. À noter que ces annotations peuvent comporter des erreurs et qu'elles émanent d'un objectif spécifique qui peut ne pas convenir au linguiste.

- Adapter ses pratiques d'accès aux données en fonction de l'outil

  L'autre solution consiste à s'adapter à l'outil pour accéder aux données et construire une méthodologie nouvelle pour établir une description linguistique des données. Cette perspective peut être perçue comme dévalorisante, puisque cela implique une forme de servitude du linguiste envers la machine, alors que sens commun voudrait que ce soit l'inverse. Pourtant les linguistes qui choisissent cette voie ne l'entendent pas de cette oreille et apportent des arguments pour justifier leur démarche.

  Dans cette approche, le concordancier est perçu comme un instrument d'exploration du corpus et le mystère règne quant à ce qui peut être découvert : personne ne peut s'y préparer, il s'agit d'affronter des armées de mots et tenter d'en faire émerger du sens, en utilisant l'expertise linguistique. Ces outils permettent d'extraire

  ➢ des listes de mots du corpus qui nous renseignent sur le type de vocabulaire employé dans un corpus donné, ainsi que sur les fréquences de ces types de formes.

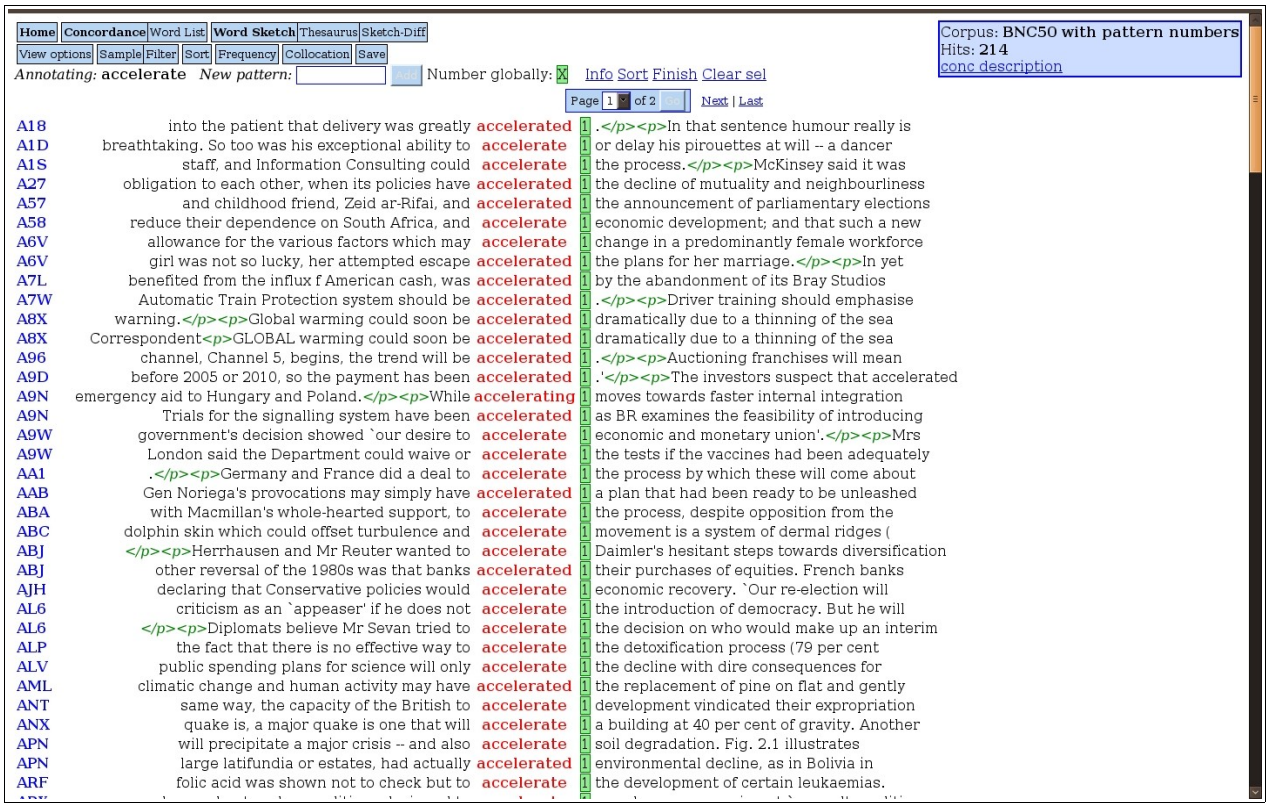


FIG 2.1 – Exemples de concordances du verbe « to accelerate » (SketchEngine)

  ➢ des concordances d'un « mot-clé », sur le format KWIC (keyword in context) qui réunit toutes les occurrences du mot cible figurant au centre de son contexte, par succession de lignes (voir figure 2.1). Ces concordances peuvent être triées alphabétiquement sur chaque position à droite ou à gauche du mot-cible (voire combiner plusieurs critères). C'est en manipulant ces différents paramètres que le linguiste cherche à identifier des régularités (des patrons, au sens de Scott & Tribble) liant le mot à son environnement immédiat, dans le texte et, en fin de compte, dans le

corpus. Un tel usage du corpus pose par conséquent avec plus d'acuité de nombreuses interrogations, concernant sa taille, son échantillonnage, sa représentativité et sa constitution [Biber, 1993]. Enfin, le fossé ainsi creusé entre théorie et pratique, nous invite à repenser la notion de système linguistique par rapport à des critères textuels comme la distance, la cooccurrence et la fréquence.

2.1.3. L'approche pilotée par le corpus

Les deux approches décrites correspondent évidemment à des abstractions : rares sont les linguistes qui observent strictement l'une d'entre elles et, de plus, d'autres pratiques sont encore développées : les concordanciers permettent l'usage de corpus étiquetés, certaines listes sont obtenues au moyen de ressources linguistiques (voir les thésaurus automatiques proposés par le SketchEngine par exemple ; [Kilgarriff et al., 2004]).

Notre tentative de classer différentes pratiques fait néanmoins écho à la dichotomie proposée par E. Tognini-Bonelli : l'approche « corpus-driven » et l'approche « corpus-based ». Comme semble l'indiquer le participe passé accolé au nom « corpus » dans ces appellations, c'est le rapport au corpus qui serait en jeu et qui les distingueraient. L'auteur explique que le participe « based » témoigne d'une relation vague entre corpus et linguistique [Tognini-Bonelli, 2001 : 65] et elle cherche à définir les principes d'une linguistique pilotée (équivalent choisi pour « driven ») par le corpus, qu'elle défend.

Le corpus devient le terrain d'investigation du linguiste (à la manière d'un sociologue qui cherche à caractériser une population donnée). Par conséquent, la validité des études est conditionnée par les critères de constitution du corpus. La constitution du corpus, étape préliminaire à l'analyse, doit pouvoir répondre aux questions suivantes : un corpus étant une collection de textes[6], quelle est la nature de la similarité de ces textes ? Quels critères employer pour les sélectionner ? La constitution de corpus est une entreprise longue et laborieuse et dans la situation actuelle, le chercheur ne peut se permettre de constituer systématiquement un corpus pour répondre à ses différentes problématiques de recherche. On assiste ainsi parfois à un renversement de situation, dans lequel le chercheur adapte ses perspectives de recherches en fonction de la disponibilité des corpus. Cette situation n'est pas nécessairement perverse, car, après tout, c'est principalement l'analyse d'instances langagières attestées qui l'intéresse, mais il doit rester conscient, au vu de ce qui a été dit, du contexte de production et de la nécessité de sa caractérisation pour s'approprier le corpus, pour pouvoir situer et interpréter ses résultats dans un cadre particulier.

Dans le cadre d'une linguistique pilotée par le corpus, la représentativité du corpus vis-à-vis du phénomène étudié devient primordiale. Le type d'analyse que l'on pourra effectuer, les résultats que l'on pourra obtenir sur un corpus donné ne seront avant tout valable que pour ce corpus. Le problème de la représentativité d'un corpus, c'est-à-dire sa capacité à constituer un échantillon de quelque chose d'autre que lui-même (d'un domaine spécifique voire de la langue générale), ne peut être pris pour acquis :

> « As we can see, when it comes to a corpus-driven approach, the issue of the representativeness of the corpus can be seen in its true importance; since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative. At present, until we know a lot more about the effect of selections on the overall picture, it is imperative to be explicit about how a corpus is constructed, and to review the relevance

6 L'approche pilotée par le corpus privilégie l'intégration de textes complets. Il est à noter que ce critère peut être substitué par l'usage d'échantillons établis sur des critères statistiques.

> of a particular corpus if there is reason to query the data it provides. […] A corpus can never be taken for granted, and must always be able to show its credentials. » [*ibid.* : 88]

Cette approche doit respecter l'intégrité des données issues du corpus :

> « The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with and reflect directly, the evidence provided by the corpus. » [*ibid.* : 84]

Une linguistique pilotée par le corpus ne correspond pas à une linguistique purement empirique (« There is no such a thing as a pure induction » ; [*ibid.* : 85]). Au contraire, elle pose avec plus d'acuité la question de la co-intégration entre données et théorie, entre quantité et signification. En effet, le mot est considéré comme indissociable du contexte dans lequel il apparaît (le corpus) : le mot n'est plus l'unité de base, c'est l'occurrence qui devient centrale. Le problème de l'association entre occurrence et unité linguistique est donc au cœur de la linguistique de corpus.

> « We must consider what kind of theory can reconcile the evidence and the statement, the instance and the system; in what way might frequency of occurrence reflect meaning, and indeed in what way should it be allowed to shape the categories and affect the system. […]
> There is, however, no escape route – corpus evidence has to be both quantified and evaluated, and part of its value lies in the quantity; the famous linguistic dichotomies have to be reviewed in the light of new data, and reconciled if possible. Statements about the system must be more directly accountable to statements about the instances. This is the stance that is taken in this study. » [*ibid.* : 49–50]

L'approche pilotée par le corpus privilégie l'usage de critères quantitatifs (la fréquence) pour faire émerger les unités et les structures. Mais elle doit également s'interroger sur la caractérisation des unités et structures au niveau de l'occurrence : l'analyse est donc contextuelle. On pourra mieux saisir cette approche en entamant une réflexion sur le contexte et c'est dans le concept de collocation qu'elle puise ses inspirations.

2.2. La collocation

J. R. Firth est la source d'inspiration première de ce « contextualisme britannique ». Il est régulièrement cité comme référence dans les ouvrages de linguistique de corpus (par exemple, [Partington, 1998 : 15] ; (Hunston & Francis 2000:230)) et est considéré comme le père du phénomène de collocation dans son acception moderne[7]. L'analyse de la collocation est indissociable de l'approche pilotée par le corpus : sa description nous permettra de comprendre pourquoi le contexte a autant d'importance.

2.2.1. La collocation comme un mode de sens

Notre traitement du concept de collocation chez Firth s'appuie principalement sur l'article « Modes of meaning », qui contient de nombreux éléments permettant de comprendre le concept et sa place au sein de sa théorie. Firth propose [Firth, 1957 : 190–215] d'introduire le terme technique de collocation qu'il considère comme un « mode de sens » parmi d'autres (phonétique, syntaxique, ou ayant trait au contexte de situation ou de culture) dans le 'spectre de description linguistique' :

> « Just as phonetic, phonological, and grammatical forms well established and habitual in any close social group provide a basis for the mutual expectancies of words and sentences at those levels, and also the sharing of these common features, so also the study of the usual collocations of a particular literary form or genre or of a particular author makes possible a clearly defined and precisely stated contribution to what I have termed the spectrum of descriptive linguistics, which handles and states meaning by dispersing it in a range of techniques working at a series of levels. » [*ibid.* : 195]

D'après cette citation, la collocation est un phénomène linguistique qui contribue à l'étude des dépendances entre mots et phrases (« the mutual expectancies of words and sentences »). Faire sens équivaut à décrire. Il est possible d'expliquer ces dépendances à différents niveaux de structure du texte :

> « The terms *structure* and *elements of structure* are not used to refer to a whole language or even to what may be called portions of a language, but exclusively to categories abstracted from common form or textual form. And quite similarly, *system, systems, terms* and *units* are restricted to a set or sets of paradigmatic relations between commutable units or terms which provide values for the elements of structure. Though structures are, so to speak, 'horizontal' while systems are 'vertical', neither are to be regarded as segments in any sense. Elements of structure, especially in grammatical relations, share a mutual expectancy in an *order* which is not merely a *sequence*. » [Firth, 1968 : 186]

Et chaque niveau contribue au sens total d'un énoncé :

> A statement of the meaning of an isolate of any of these [languages] cannot be achieved at one fell swoop by one analysis at one level. [Firth, 1957 : 192]
> The question, therefore, is not how much meaning can be excluded, but how much meaning can legitimately be included. It might even be said that meaning *must* be included as a fundamental assumption. [Firth, 1968 : 50]

7 Antérieurement, H. E. Palmer [Palmer, 1933] a également employé ce terme.

2.2.La collocation

La notion de collocation peut se confondre avec celle de dépendance mutuelle (« mutual expectancies ») : l'auteur emploie par exemple le terme de collocation grammaticale pour désigner des dépendances à ce niveau (voir « grammatical collocation » [Firth, 1957 : 197] et « word collocation » [*ibid.* : 198]. Par conséquent,

- Soit la collocation est perçue comme un outil : elle intervient à tous les niveaux d'analyse linguistique, auquel cas la collocation désigne uniquement une relation de co-occurrence.
- Soit la collocation est un phénomène qui relève du niveau du mot : elle n'intervient qu'au niveau de co-occurrence de mots, et elle est reconnue comme un niveau propre. Si l'on souhaite appliquer ce concept à d'autres niveaux d'analyse, d'autres termes permettront de les distinguer. C'est ce qu'il semble faire en employant le terme de « colligation » pour le niveau syntaxique [*ibid.* : 186].

Comme nous allons le démontrer, il s'agit de la seconde possibilité. Firth définit la collocation par exclusion. Elle n'est pas une association d'idées comme on pourrait l'interpréter dans une sémantique conceptuelle (au moyen d'un méta-langage, ou d'un discours sur le discours ; « by means of further sentences in shifted terms ») :

> « The statement of meaning by collocation and various collocabilities does not involve the definition of word-meaning by means of further sentences in shifted terms. Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. » [*ibid.* : 196]

Remarquons que ce n'est pas parce que la collocation n'implique pas « d'association d'idées », qu'il évacue cette dernière du prisme de description linguistique (comme on a pu le dire du structuralisme de Bloomfield ou de Harris ; [Bloomfield, 1933], [Harris, 1951]) :

> « In doing this, I must not be taken to exclude the concept of mind, or to imply an embracing materialism to avoid a foolish bogey of mentalism. » [Firth, 1957 : 192]

Le sens collocatif ne correspond pas non plus au sens qu'il aurait en contexte dans une 'sémantique située', qui lierait les associations par collocation à des associations d'entités dans ledit 'contexte de situation' (situation d'énonciation) :

> « It must be pointed out that meaning by collocation is not at all the same thing as contextual meaning, which is the functional relation of the sentence to the process of a context of situation in the context of culture. » [*ibid.* : 195]

Firth veut attirer l'attention des linguistes sur le mot. Ce qui caractérise la collocation est, non pas le contexte et sa nature, mais l'unité dont on cherche à caractériser le contexte, le mot (par opposition à la catégorie grammaticale). La collocation est le niveau d'analyse du mot, comme le confirme la citation suivante :

> « In this connection, I would like to put forward the concept of *collocation* which I have introduced in my own work. This is the study of key-words, pivotal words, leading words, by presenting them in the company they usually keep – that is to say, an element of their meaning is indicated when their habitual word accompaniments are shown. The collocations presented should usually be complete sentences and, if it is conversation, the collocations should be extended to the utterances of preceding and following speakers. » [Firth, 1968 : 106–107]

Le niveau collocationnel du sens est de l'ordre de la phraséologie, il concerne les associations habituelles de mots sans les interpréter ; il se limite à leur description et à leur corrélation à des niveaux de langues ou à des types de texte :

> « One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night. This kind of mutuality may be paralleled in most languages and has resulted in similarities of poetic diction in literatures sharing some classical sources. » [Firth, 1957 : 196]

Remarquons que l'environnement collocationnel n'est pas nécessairement un mot, comme on pourrait le croire avec l'exemple *dark night*, mais peut s'étendre à la phrase (« commonest sentences ») ou à des expressions. Par exemple, les collocations qu'identifie Firth pour le mot ass sont *Don't be an* _ ou *You silly _ !* [*ibid.*]. Les relations collocationnelles peuvent ainsi s'étendre à des relations 1/n mots.

En dehors de ce qu'ils peuvent signifier ou de ce à quoi ils peuvent faire référence, les mots auraient une véritable identité physionomique[8], une empreinte (graphique ou sonore), qui se découvre en contexte (voir notamment [De Beaugrande, 1991] et [Kachru, 1980]; [Léon, 2007] [Léon, 2008]) :

> « Words stare you in the face from the text, and that is enough; and as Wittgenstein said, a word in company may be said to have a physiognomy. The elements of style can be stated in linguistic terms. » [Firth, 1957 : xii (introduction)]

Pour poursuivre cette personnification, les mots « vivent » en groupe, toujours en « compagnie », certains s'attirent, d'autres se repoussent. Il serait ainsi plus approprié de traduire le verbe *to know* par *reconnaître* (car connaître, c'est reconnaître) dans le célèbre adage firthien[9] :

> « You shall know a word by the company it keeps! » [Firth, 1968 : 179]

Une seconde interprétation peut enrichir la notion de collocation : cette dernière désigne les contextes « communs » dans lesquels apparaît un mot, à l'opposé de contextes singuliers, incongrus ou obsolètes. Elles sont ordinaires, « usuelles » dans le sens où elles traduisent des usages. L'auteur propose de distinguer l'usage général de l'usage technique ou personnel de la collocation :

> « This kind of study of the distribution of common words may be classified into general or usual collocations and more restricted technical or personal collocations. The commonest sentences in which the words *horse, cow, pig, swine, dog* are used with adjectives in nominal phrases, and also with verbs in the simple present, indicate characteristic distributions in collocability which may be regarded as a level of meaning in describing the English of any particular social group or indeed of one person. » [Firth, 1957 : 195]

8 La citation de L. Wittgenstein est la suivante :

> The meaning of a word is not the experience one has in hearing or saying it, and the sense of a sentence is not a complex of such experiences.[…] The sentence is composed of the words, and that is enough. Though-one would like to say-every word has a different character in different contexts, at the same time there is *one* character it always has: a single physiognomy. It looks at us.-But a face in a painting looks at us too. [Wittgenstein, 1953 : 181]

9 Vu ce qui a été dit précédemment, ne traduisons pas *connaître* par *connaître la signification d'un mot*. De la même manière, la collocation n'épuise pas non plus la connaissance d'un mot.

Comme exemple d'application de la collocation usuelle restreinte à une personne (un style), il propose de caractériser la spécificité stylistique des poèmes de Swinburne (le « Swinburnais ») :

> « From the point of view of linguistic criticism there is sufficient evidence to show that much of the Swinburnese vocabulary, embedded in his typical collocations with their prosodies, takes its form from his patterns of opposition, requiring such phrases as 'Mis-trust and trust'. » [*ibid.* : 200]

Enfin, comme exemple d'application de la collocation restreinte à un domaine technique, il procède à une courte analyse « diachronique » d'écrits épistolaires des 18e et 19e siècle pour caractériser l'anglais correct de l'aristocratie (« King's English ») :

> « A linguistic study of the letters of the upper class seems to show persistent features of what we call the King's English in modern times, and these features have been shared by increasing numbers of writers and speakers in the nineteenth century and up to the present time, largely perhaps as a result of the influence of the big public schools, the older universities, and the snob value of the aristocratic. » [*ibid.* : 206]

En résumé, la collocation est un concept protéiforme qui peut inclure les notions suivantes :

- la dépendance
- le mot
- le contexte
- la phraséologie (l'identité physionomique)
- l'usage

De nombreuses questions restent en suspens, comme celle des critères pour les caractériser.

Dans un ouvrage dédié à la mémoire de Firth (publié en 1966), M. A. K. Halliday et J. McH. Sinclair introduisent les problématiques et méthodes liées au niveau de collocation ou niveau du mot, qu'ils dénomment *Lexis*. Tous d'eux s'approprient le concept de collocation mais leurs approches divergent : Halliday, qui a réalisé sa thèse sous la direction de Firth, se positionne dans la continuité de ses travaux en considérant la lexie (la collocation) comme un niveau de description à part entière et en cherchant à l'articuler avec d'autres niveaux de description, grammatical et « sémantique ». Sinclair, pour sa part, saisit l'opportunité de la singularité du phénomène de collocation pour proposer des pistes de recherche. Il faut remarquer que l'approche développée par Sinclair dans cet article diffère de celle qu'il développera dans ses publications ultérieures, et pour lesquelles il est reconnu ([Sinclair, 1991] par exemple), lorsqu'il adoptera une perspective lexicographique et non plus firthienne du sens.

2.2.2. La Lexie comme niveau linguistique

Halliday s'attaque au problème de la conception d'un niveau lexical qui soit distinct du niveau grammatical et tout à la fois linguistique, c'est-à-dire complémentaire au niveau grammatical. En ce sens, les outils conceptuels sont d'un genre différent (« as a different kind » ; [Halliday, 1966 : 148]), mais permettent tout autant de contribuer à la description d'unités textuelles. Il établit le parallèle entre ces deux niveaux en indiquant qu'une séquence de cinq « items » peut constituer cinq unités grammaticales et cinq unités sur le plan lexical :

> « So in *a strong cup of tea* the grammar recognizes (leaving aside its higher rank status, for example as a single formal item expounding the unit 'group') five items of rank 'word' assignable to classes, which in turn expound elements in structures and terms in systems; and the lexis recognizes potentially five lexical items assignable to sets. » [*ibid.* : 155]

Le niveau lexical est appréhendé d'un point de vue systémique (c'est un système) et se situe ainsi à un niveau d'abstraction détaché du texte.

Pour Halliday, les méthodes qui permettent l'identification de faits lexicaux ne reposent pas nécessairement sur des critères statistiques, ni non plus sur des observations en corpus [*ibid.* : 149–150, note 6]. L'auteur souhaite distinguer clairement lexique et corpus, l'observation des usages et leur quantification constituant des moyens pour la description du lexique.

L'auteur distingue trois types de structures possibles pour le lexique : l'item, l'ensemble et la collocation [*ibid.* : 152]. Ces notions se définissent mutuellement [*ibid.* : 153], ce qui implique une circularité :

> « There is of course no procedural priority as between the identification of the paradigmatic and syntagmatic relations into which they enter: 'item', 'set' and 'collocation' are mutually defining. » [*ibid.* : 150]

Les items sont associés sur le plan paradigmatique dans des « ensembles » (« sets »[10]), en fonction de leur similarité collocationelle, c'est-à-dire en fonction de leur association sur le plan syntagmatique :

> « If we say that the criterion for the assignment of items to sets is collocational, this means to say that items showing a certain degree of likeness in their collocational patterning are assigned to the same set. » [*ibid.* : 158]

La collocation est le critère qui permet d'associer des items dans des ensembles sur le plan lexical.

Une telle formalisation paraît plus simple qu'elle est en vérité, car elle masque un certain nombre d'opérations que nous allons à présent déplier. Trois points nécessitent d'être éclaircis.

(1) La collocation et l'item ne peuvent pas être des occurrences, bien qu'Halliday nous dise qu'ils appartiennent à l'axe syntagmatique. En effet, si un item était une occurrence, on ne pourrait, en toute logique, associer dans des ensembles, que des occurrences. Mais travailler au niveau de l'occurrence n'a aucun sens, du moins si l'on cherche à constituer des ensembles : chaque occurrence est unique donc chaque collocation lierait deux items qui seraient tous deux uniques. On ne pourrait donc pas établir de comparaison entre *cup* dans un énoncé comme (prononcé à un temps *t*) *a strong cup of tea*, et *cup* dans une énoncé (prononcé à un temps *t+1*) comme *a strong cup of tea* parce que malgré l'identité exacte des sons ou des formes, ce sont deux occurrences différentes ! Le même raisonnement vaut pour la constitution d'ensemble ou de collocations. Aussi incongru que cela puisse paraître, il est nécessaire de distinguer au moins deux niveaux d'analyse afin d'envisager la comparaison d'unités et rendre un intérêt à la notion de collocation. La relation de collocation concerne donc, par nécessité, des types (déjà des catégories) dont les occurrences sont associées sur le plan syntagmatique.

10 Pour Halliday, l'ensemble est l'équivalent du terme « système » en grammaire sur le plan lexical

(2) Le type qui nous vient automatiquement à l'esprit est la forme, qui, avant d'être caractérisée en termes collocationnels, est définie en fonction de l'identité de forme graphique ou sonore. Alors, dans ce cas, *cup* dans *a strong cup of tea* est le même item que *cup* dans *a strong cup of tea* et il y a un sens à parler de collocation entre types <*cup, tea*> pour lesquels on aurait identifié un certain nombre d'occurrences et à constituer par exemple un premier ensemble minimal *{cup}* sur la base de la collocation <*cup, tea*>.

(3) À ce stade, l'analyse collocationnelle est opérationnelle mais a peu d'intérêt si l'on s'arrête là. En effet un ensemble équivaut uniquement à une collocation et il y a autant d'ensembles pour un même type-forme que de collocations entre type-formes : l'ensemble *{cup$_1$}* issu de la collocation <*cup, tea*>, l'ensemble *{cup$_2$}* issu de la collocation <*cup, of*>, et ainsi de suite. Pour envisager la caractérisation de la forme-type *cup* dont les occurrences apparaissent dans des environnements collocationnels différents, il faut autoriser la comparabilité et la fusion des ensembles. La complexité de cette opération de fusion est proportionnelle au nombre de collocations identifiées. Si donc, un type-forme est caractérisé par plus d'une collocation, c'est que les ensembles sont considérés équivalents vis-à-vis de ce type-forme et qu'il est utile de distinguer au moins un niveau intermédiaire, que l'on pourra nommer un ensemble complexe, entre l'ensemble (simple) et le type-forme.

Ce parcours nécessaire d'une analyse collocationnelle permet de mieux saisir la complexité des objets que l'on manipule dans un texte. Nous n'avons pas évoqué les dimensions supplémentaires de taille du co-texte (quelle distance respecter entre deux occurrences pour les lier par collocation) ni d'unité (quelle taille définir pour les éléments liés par collocation), mais ce sont différentes réalisation du même schéma fondamental (tableau 2.1).

Types	[strong]
Ensembles Complexes	{{strong-tea}-{strong-man}}
Ensembles	{strong-tea} {strong-man}
Occurrences	… strong tea … strong tea … strong man …

Tableau 2.1 – De l'occurrence au type

La nature du type (catégorie), que nous avons utilisée pour illustrer notre raisonnement, était la forme, mais il est tout à fait possible de proposer d'autres types établis sur d'autres critères que la simple identité de forme graphique (par exemple la forme canonique, le référent, le concept, etc.). Enfin, il faut également envisager la possibilité d'établir des relations de collocations entre des unités de type différent.

On peut identifier quatre relations entre occurrence et type, réunis figure (2.2) :

- Comme nous l'avons vu, le candidat au statut de type lexical peut correspondre en premier lieu à la forme (« type-token » [*ibid.* : 157] ; figure 2.2, schéma 1a).
- Comme le remarque Halliday, des occurrences différentes d'une même forme ne sont pas nécessairement associées au même type ([*ibid.* : 153] ; figure 2.2, schéma 2a). Par exemple, la forme *ride* en anglais sera tantôt un verbe, tantôt un nom, soit deux types différents.

- On peut également être amené à associer deux formes différentes à un même type, ce que propose Halliday pour les formes *strong* et *strongly*, qui relèvent d'après lui du même type (figure 2.2, schéma 3a). Cette complexité sur le plan lexical veut traduire les similarités collocationnelles de plusieurs formes :

  « What is abstracted is an item *strong*, having the scatter *strong, strongly, strength, strengthened*, which collocates with items *argue* (*argument*) and *tea*; and an item *power* (*powerful*, *powerfully*) which collocates with *argue* and *car*. » [*ibid.* : 151]

- Enfin, Halliday indique que les types peuvent correspondre à une séquence de formes comme les noms composés (*pomme de terre*) (figure 2.2, schéma 4a).

Étant donné qu'une forme peut renvoyer vers plusieurs types (phénomène que l'on peut appeler *polycatégorialité*) et qu'entre autres, plusieurs formes peuvent renvoyer vers un même type, nous avons représenté ces quatre relations en déclinant les trois niveaux : celui de l'occurrence, de la forme et du type (figure 2.2 ; schémas 1b, 2b, 3b et 4b).

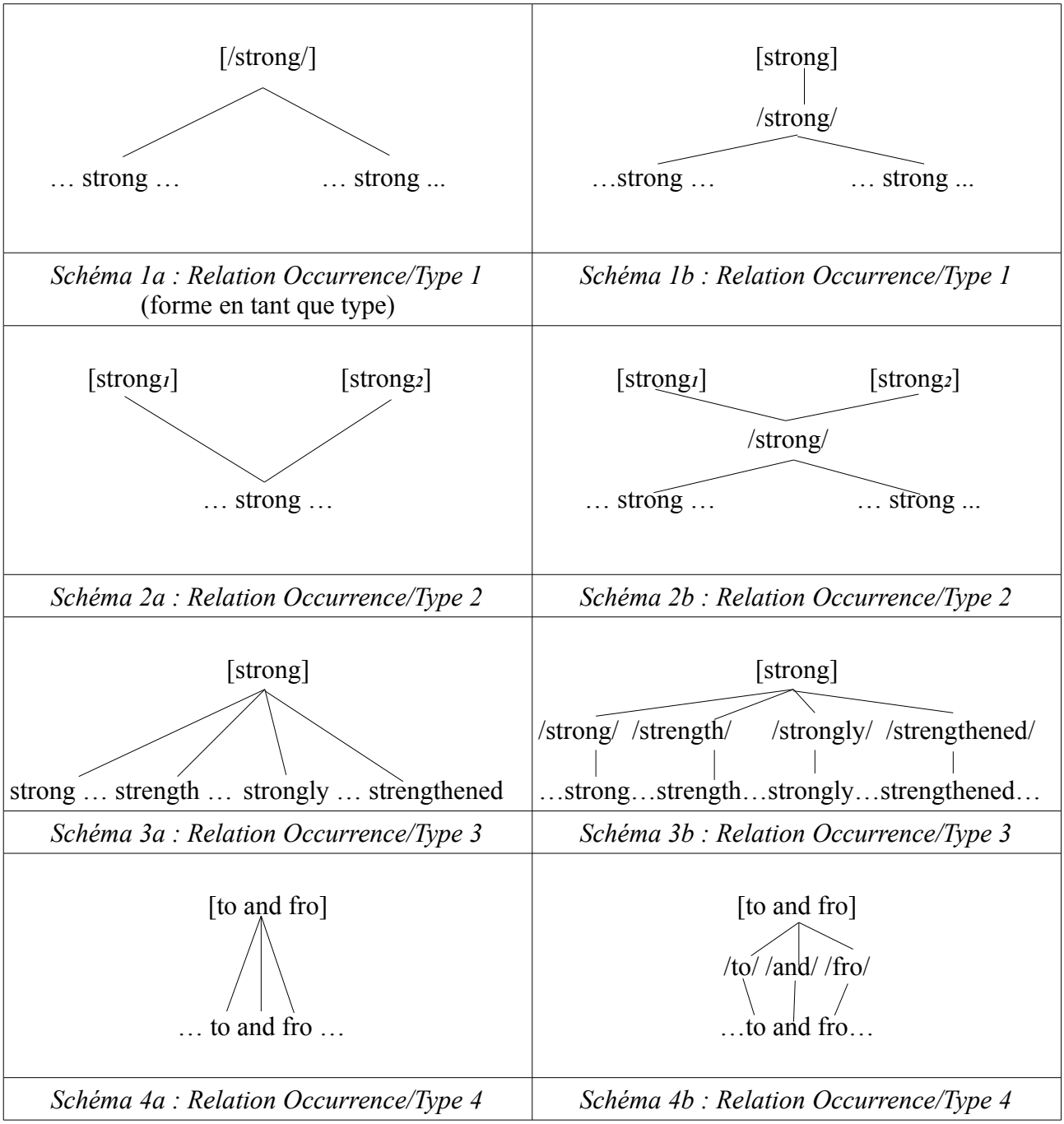


Schéma 1a : Relation Occurrence/Type 1 (forme en tant que type)

Schéma 1b : Relation Occurrence/Type 1

Schéma 2a : Relation Occurrence/Type 2

Schéma 2b : Relation Occurrence/Type 2

Schéma 3a : Relation Occurrence/Type 3

Schéma 3b : Relation Occurrence/Type 3

Schéma 4a : Relation Occurrence/Type 4

Schéma 4b : Relation Occurrence/Type 4

FIG 2.2 – Schémas de relations entre Occurrence, Forme et Type
les pointillés signifient des espaces de textes (occurrences), les formes sont entre barres obliques et les types entre crochets

Dans sa tentative de formalisation du niveau lexical (que nous avons développée), Halliday a cherché à le démarquer du niveau grammatical, tout en l'analysant sur un même schéma global. Bien que ce ne soit qu'une première approche, elle permet d'établir quelques principes pour l'analyse de la lexie, sur des critères qui lui sont propres. Des questions restent en suspens, comme celle du rapport entre catégories grammaticales et lexicales et entre les liens qui peuvent exister sur ces deux plans :

> « It is not known how far collocational patterns are dependent on the structural relations into which the items enter. For example, if *a cosy discussion* is unlikely, by comparison with *a cosy chat* and *a friendly discussion*, is it the simple co-occurrence of the two items that is unlikely, or their occurrence in this particular structure? All that has been said above has implied an approach in which grammatical restrictions are not taken into account, and reasons have been given for the suggestion that certain aspects of linguistic patterning will only emerge from a study of this kind. But it is essential also to examine collocational patterns in their grammatical environments, and to compare the descriptions given by the two methods, lexical and lexicogrammatical. » [*ibid.* : 159]

Nous savons que le statut même de forme (comme nature de type particulière) repose sur des critères orthographiques, du moins des critères qui ne sont pas lexicaux. Pour poursuivre l'analogie type-occurrence, les catégories grammaticales peuvent être considérées comme des types. Le lien entre relations « structurales » (comprendre « grammaticales » dans la terminologie de Halliday, comme <*Nom, Adjectif*>) et relations collocationnelles consistera ainsi en l'étude des corrélations entre des types de nature différente. Halliday ne pousse pas jusqu'à envisager de telles corrélations sur des critères « sémantiques » (corrélations entre relations collocationnelles et relations sémantiques), mais insiste pour que le niveau lexical reste autonome et distinct du niveau sémantique au sens classique du terme [*ibid.* : 159–160].

2.2.3. Les procédures de découverte d'unités lexicales

Sinclair entame son article en adoptant une position que l'on peut désormais qualifier de firthienne, qui consiste à conserver l'autonomie du niveau lexical (« language form ») pour en identifier les usages (« patterns »), tout en indiquant qu'il s'agit surtout de se libérer autant que possible d'*a priori* :

> « When we are studying language form, we are concerned to examine the way in which patterns recur, without taking into account other patterns which lie outside language, patterns of social or natural organization. It would be surprising indeed if a great many of these external patterns were not reflected in language, but we try not to inflate this expectation into preconception. » [Sinclair, 1966 : 410]

La « préconception » est un concept récurrent dans les travaux de Sinclair qui traduit son souci de la validation empirique des hypothèses et sa volonté de respecter l'intégrité des données. Le sens « révélé » par les ensembles de collocations (qu'il nomme des « clusters ») est dit « formel » :

> « Firth said (op. cit., page 196) 'One of the meanings of *night* is its collocability with *dark*' and we can go on from there to say that the formal meaning of an item A is that it has a strong tendency to occur nearby items B, C, D, less strong with items E, F, slight with G, H, I, and none at all with any other item. And that is exactly what is tabulated in the cluster. » [*ibid.* : 417]

Sinclair se propose d'aborder le problème du statut d'une unité lexicale (sa taille en mots). Cette unité (le type) n'est pas « donnée » [*ibid.* : 412], il propose de la découvrir grâce à la collocation. Mais la collocation d'une unité s'établit avec d'autres unités : sa définition est donc circulaire, ce qu'il ne manque pas de noter [*ibid.*]. L'auteur se demande s'il est possible d'éviter de s'appuyer sur des critères externes (comme l'identité de forme ou « mot orthographique » ; [*ibid.*]) et indique que seule une étude de textes longs pourra y répondre. Il envisage donc trois points de

départ à son analyse linguistique [*ibid.* : 419–420] :

- le mot orthographique (la forme) : il a l'inconvénient de disperser des occurrences d'une même unité lexicale (*strong*, *strongly* par exemple).
- l'annotation préalable du linguiste : il la considère impraticable et peu fiable.
- l'unité minimale, qu'il nomme le morphème : c'est de son point de vue la plus fiable.

En considérant le morphème comme unité de départ, il reste deux phénomènes à prendre en compte pour affiner la notion d'unité lexicale :

> « At least we could be confident that no lexical item was smaller than the chosen unit, and the remaining problems would be:
> (i) detecting multi-morpheme items;
> (ii) detecting more than one item with the same form. » [*ibid.* : 420]

Détaillons ces deux points :

> (i) Pour la détection d'unités complexes (les verbes à particule, les locutions figées, les proverbes, etc.), l'intuition de Sinclair est la suivante : les collocations d'une telle unité correspondront à l'intersection des collocations de chaque unité qui la compose. Plus précisément les unités qui la composent seront toutes associées aussi significativement (en fonction de la fréquence) aux mêmes collocations [*ibid.* : 422–423].
>
> (ii) Concernant l'analyse du second phénomène, qui consiste à détecter plusieurs unités à partir d'une même morphème (*ride* en tant que nom ou en tant que verbe), Sinclair propose la notion d'inter-collocation [*ibid.* : 425]. Si deux collocations d'une unité ne sont pas l'une et l'autre en collocation, alors cette unité peut être divisée. Pour prendre un exemple en français, on cherche à savoir si la forme *livre* correspond à une ou plusieurs unités. *Livre* est en collocation avec, d'une part, *page*, *reliure* et d'autre part *magasin*, *domicile*. Si *page* et *magasin* ne sont pas en collocation, si *reliure* et *domicile* non plus (ainsi que les autres combinaisons possibles), c'est qu'ils constituent des environnements différents (appartenant à des réseaux collocationnels différents) et que nous sommes peut-être en présence de deux unités de *livre* (le nom et le verbe).

Pour tester ces hypothèses, l'auteur propose de définir les termes d'une analyse collocationnelle [*ibid.* : 415] : le nœud (« node ») est le mot-cible que l'on cherche à analyser, l'empan (« span ») est l'espace de texte compté en nombre de mots à gauche et à droite autour du nœud (pour le terme « span »), et les collocats (« collocates ») sont les unités apparaissant dans cet empan vis-à-vis du nœud. La taille de l'empan, comme la distance/proximité d'un nœud et de son collocat ne peut être déterminé par avance [*ibid.*]. Il propose une analyse d'unités lexicales dont le statut semble « évident » (« fairly obvious », [*ibid.*]) d'un court texte en choisissant un empan de 3 unités à droite et à gauche du nœud. Les collocats sont comptés, ordonnés et regroupés dans une liste d'environnement (« total environment table »), en fonction de leur fréquence d'apparition dans l'empan [*ibid.* : 416]. C'est de ce tableau que sont censés émerger les « clusters ». Un cluster désigne l'ensemble des collocations « significatives » d'un nœud donné : elles correspondent par exemple à la collocation proposée par Firth entre *dark* et *night*. Pour les isoler, Sinclair cherche à savoir dans quelle mesure un nœud est prédit par une collocation :

> « To summarize, we measure both the way in which an item predicts the occurrences of others (and get results such as the table on page 416), and also the way in which it is predicted by others. Then we choose the second of these measurements for our statements of the lexical meaning of the item. This statement is called a cluster and is derived from the total environment tables. » [*ibid.* : 417]

Mais comme il le remarque, l'identification de ces clusters ne peut se faire automatiquement à partir d'une liste de collocats, car le critère de fréquence d'association n'est pas une mesure pertinente de la contrainte exercée par le nœud. Il faut également tenir compte de la fréquence absolue du collocat. Certains collocats comme *save* pour le nœud *money* ont une fréquence en collocation égale à leur fréquence absolue, alors que d'autres collocats comme *think* ont une plus forte fréquence absolue. Ce dernier mot doit être moins « significatif », puisqu'il apparaît en collocation avec d'autres nœuds. Les morphèmes dits grammaticaux (*le*, *est*, *de*, en français par exemple) en sont un exemple plus explicite [*ibid.* : 422–423].

Une collocation qui n'a de pouvoir prédictif sur le nœud est dite casuelle, et c'est la plus représentée [*ibid.* : 418]. Elle n'est pas nécessairement habituelle ou régulière, mais peut être accidentelle, atypique. Une collocation significative concerne deux unités répétitivement associées. Remarquons que cette définition quantitative de la collocation correspond à une conception quantitative des données, dans laquelle notre capacité à distinguer le typique de l'atypique (significatif) est proportionnel au nombre d'instances que l'on aura pu obtenir.

Pour les identifier, Sinclair propose une mesure de significativité qui traduise la probabilité de co-occurrence de deux unités dans un texte [*ibid.*]. Soit p le nombre d'unités total d'un texte, s l'empan choisi (le nombre de mots à droite et à gauche), f la fréquence absolue du nœud et n la fréquence absolue de son collocat, la probabilité de co-occurence de deux unités serait $\frac{nsf}{p}$:

> « Then if we consider a particular node which occurs n times in the text, the probability of our item collocating with this node is $\frac{nsf}{p}$. » [*ibid.*]

Malheureusement, cette mesure ne correspond pas à ce qu'il recherche : elle ne calcule pas l'attractivité de deux unités et ne permet pas non plus d'isoler les collocations dites « significatives ». Tout d'abord, cette mesure ne distingue pas la part de contrainte de chaque unité de la collocation, puisque $\frac{nsf}{p}$ et $\frac{fsn}{p}$ donneront le même résultat. Comparer la valeur de cette mesure avec la fréquence de cooccurrence des deux unités en question, comme il le suggère, ne le permettra pas non plus, puisque la fréquence de cooccurrence sera identique pour ces deux unités. Ensuite, si cette mesure est employée à des fins de comparabilité entre collocats pour isoler ceux qui sont significatifs, elle n'ajoute rien au critère de fréquence, puisqu'elle augmente en fonction des fréquences respectives du nœud et du collocat et ne prend donc pas en compte la contrainte de l'empan ; si la fréquence d'une unité est forte, sa fréquence relative sera forte, et la valeur de sa mesure sera forte : autant choisir la fréquence.

Les mesures qui seront employées par la suite en linguistique de corpus (le z-score et l'information mutuelle, notamment) permettront de mieux caractériser cette notion de « collocation significative », car elles établissent la significativité d'une collocation en fonction de son caractère inhabituel, ni trop ni trop peu fréquent, dans une configuration donnée (par rapport à leurs distributions respectives).

2.3. Lexique et sens : les travaux de Sinclair

Les travaux présentés dans la partie précédente se définissent particulièrement par la considération de l'autonomie de la lexie. Les contributions de Sinclair que nous allons décrire dans cette partie va permettre de fédérer une communauté autour de l'analyse linguistique de corpus numériques, la linguistique de corpus. Si l'analyse des formes restera une préoccupation constante, Sinclair va progressivement inclure dans sa réflexion le problème du sens.

2.3.1. Les résultats du rapport OSTI

La collocation est le thème principal du rapport OSTI [Sinclair et al., 2004] qui fait un bilan des recherches menées en collaboration avec un statisticien. Deux problématiques de recherche y sont explorées, la possibilité d'introduire une mesure de collocation et les corrélations entre collocation et sens :

> « (a) how can collocation be objectively described? (b) what is the relationship between the physical evidence of collocation and the psychological sensation of meaning? » [*ibid.* : 3]

Sinclair reprend particulièrement les deux problèmes posés par l'identification d'une unité lexicale (cf. *infra* 2.2.3). Le premier concerne l'utilisation de la collocation (l'inter-collocation ; cf. *infra* 2.2.3) pour distinguer deux sens d'une même forme. D'après l'auteur, les résultats (sur un corpus de 1,2 millions de mots) montrent que la collocation possède un fort potentiel de désambiguïsation, bien que les mots soient rarement polysémiques dans le corpus :

> « It was assumed that a word which realized two entirely separate meanings would show two separable groups of collocates, so the program measured the intercollocation among all the collocates of selected nodes. The technique was modified greatly during development, and yielded the following results: (a) the technique is remarkably sensitive. Even with the limitations of text size and machine flexibility, and using only collocational information at a span of +-2 or +-4, a large proportion of instances of ambigous words were correctly assigned (b) the adding of positional information about collocates improved results in some cases (particularly where different meanings of a word correlated with grammatical categories) (c) text produced by informants being asked to write specimen sentences around selected words gave very good results (d) it is not easy to find sufficient examples in a single text of each meaning of an ambiguous word. One of the meanings nearly always preponderated. » [*ibid.* : 5–6]

Le second problème concerne la détection d'unités complexes. Elles portent dorénavant le nom d'« idiome », qui est redéfini comme une unité de sens fédérant plusieurs mots (« unique semantic partnership »). Le programme informatique conçu à leur détection a visiblement échoué :

> « Lexicography has also traditionally recognized that certain strings of words form unique semantic partnership, called idioms, constituting further argument against the word as the unit of lexical patterning. A technique was devised for assessing the cumulative significance of successive words in a text, but was abandonned because repeated experimentation failed to reduce the arbitrariness of many of the programming decisions. » [*ibid.* : 6]

Le critère de définition d'une unité lexicale doit être sémantique : le niveau lexical n'est pas autonome :

> « A lexical item is a unit of language representing a particular area of meaning which has a unique pattern of co-occurrence with other lexical items. » [*ibid.* : 9]

Si (selon le principe de circularité encore à l'œuvre ici ; cf. *infra* 2.2.2 et 2.2.3) une unité lexicale est définie comme une unité de sens, alors elle se définit par sa co-occurrence avec des unités de sens. Ce changement de perspective a une incidence importante, car cela suppose que les unités de sens peuvent être délimités, que la combinaison de ces unités sémantiques obéit à une systématique, enfin, que la lexie n'est plus indépendante. La forme et le sens sont liés, une modification de la forme implique une modification du sens et vice et versa :

> « The recognition that form is often in alignment with meaning was an important step, and one that cut across the received orthodoxy of the explanation of meaning. Soon it was realized that form could actually be a determiner of meaning, and a causal connection was postulated, inviting arguments from form to meaning. Then a conceptual adjustment was made, with the realization that the choice of a meaning, anywhere in a text, must have a profound effect on the surrounding choices. It would be futile to imagine otherwise. There is ultimately no distinction between form and meaning. » [Sinclair, 1991 : 7]

Ce revirement dans la direction des recherches sera développé dans un recueil d'articles (entre 1985 et 1990) recomposé sous forme de livre, qui est devenu en quelque sorte la « bible » de la linguistique de corpus britannique parce qu'il pose les bases de l'analyse sémantique dans le cadre d'une linguistique de corpus informatisée. Trois pistes principales sont développés dans cet ouvrage [Sinclair, 1991] : le rapport entre sens et structure, le principe d'idiome et les unités étendues de sens.

2.3.2. Sens et structure

Le projet lexicographique de dictionnaire COBUILD (Collins Birmingham University International Language Database) avait pour objectif de créer une nouvelle génération de dictionnaire qui s'appuie sur les résultats d'une analyse d'un large corpus (le Birmingham Corpus, de 20 millions de mots en 1987). Les études menées sur ce corpus ont notamment mis en valeur la relation entre sens et structure. Les patrons grammaticaux d'un mot étaient fortement associés à un des sens de ce mot (et vice et versa) :

> « In nearly every case, a structural pattern seemed to be associated with a sense. Despite the broad range of material in the corpus, when instances were sorted into ‘senses’, a recurrent pattern emerged. » [Sinclair, 1987 : 109]
> « There was in practice no clear distinction between grammar and lexis, and grammatical rules merged with restrictions in particular instances, and those restrictions ranged from the obviously grammatical to the obviously lexical. » [*ibid.* : 110]

En plus de corrélations entre grammaire et lexie, un des apports de l'observation en corpus est que la forme même du mot peut être employé comme critère dans la détermination de son sens. Avant même de considérer un mot dans sa forme canonique (lemme), un certain nombre d'informations peuvent être obtenues sur chaque forme :

> « There is a lot to be learnt about language from the study of it in this simple format. Most studies leap ahead and group the crude words according to simple notions of meaning, instead of deriving as much information as possible from each stage in the developing sophistication of description. » [Sinclair, 1991 : 41]

Le processus de lemmatisation est pointé du doigt car il implique une catégorisation de sens préalable (en réunissant plusieurs formes sous un même type) :

> « Lemmatization looks fairly straightforward, but is actually a matter of subjective judgement by the research. There are thousands of decisions to be taken. Also, it is not yet understood how meanings are distributed among forms of a lemma, and a new branch of study is looming – the interrelationships of a lemma and its forms. » [*ibid.*]

Sinclair pense que cette catégorisation devrait plutôt s'appuyer sur l'usage des formes qu'un lemme sous-tend. Il illustre ce problème en étudiant les concordances du verbe *to decline*.

- L'analyse de corpus peut tout d'abord nous renseigner sur la distribution de la fréquence de chaque forme de ce verbe [*ibid.* : 44]. En les comparant avec celles que peut fournir un dictionnaire, l'auteur constate que certaines sont absentes et que d'autres sont très fréquentes (50% pour la forme *decline*).
- Pour chaque forme, il analyse leur distribution grammaticale (*decline* comme verbe ou nom) : il observe que la forme *decline* est majoritairement employée en tant que verbe dans son corpus (88.5%), mais aussi que l'usage verbal prédomine pour les autres formes (99% ; [*ibid.* : 45, fig. 2]).
- Les patrons grammaticaux ne sont pas tous référencés dans la dictionnaire, comme il le constate pour l'usage adjectival de *declining*.
- Les sens d'un lemme peuvent être quantifiés en prenant appui sur les définitions du dictionnaire. Pour chaque forme, il cherche à lui associer une définition. Par exemple, la forme *declined* est fortement employée dans le sens « to refuse to do or accept (something), esp. politely » [*ibid.* : 46]. Il observe également que certaines définitions sont très proches et mériteraient d'être associées et enfin que d'autres définitions peuvent être révélées par une analyse de corpus.

Cette première approche n'est pas une analyse collocationnelle[11] et s'appuie sur le dictionnaire pour définir le sens. La définition de certaines formes, dont la fréquence est forte, comme *of*, ou dont le « mot orthographique » (séquence de caractères alphanumériques entre deux espaces) est associé à des formules idiomatiques, comme *set* (*set in*, *set up*, *set off*, etc.), nécessite la prise en compte du contexte :

> « 'What does set mean?' is hardly a sensible question. It has to be put into context, because in most of its usage it contributes to meaning in combination with other words. » [*ibid.* : 67]

Pour caractériser le contexte d'un mot, il propose par exemple sa position dans une séquence canonique comme la proposition ou la phrase et la taille de ces séquences. Il observe par exemple que la locution verbale *set in* apparaît généralement dans de courtes phrases [*ibid.* : 74], dans des

11 Sinclair observe et trie seulement les concordances d'un nœud (les différentes formes du lemme)

propositions subordonnées, et en position finale[12]. L'analyse du contexte peut également s'appuyer sur la nature sémantique des mots liés par une relation syntaxique. Par exemple, *set in* (intransitif) se combine avec des sujets désignant une situation désagréable [Sinclair, 1991 : 75], comme *rot*, *decay*, *malaise*, *despair*, etc.

> « The distinguishing criteria are commonplace features of grammar or semantics, and even in the small group of phrasal verbs with *set*, we can see them beginning to recur. » [*ibid.* : 78]

Sinclair propose donc de mettre à profit toutes les ressources à la disposition du linguiste pour caractériser une unité afin d'en obtenir une vision synthétique [*ibid.* : 154–156, tableaux 1 et 2 de l'annexe III] qui, d'une certaine manière renoue avec la vision systémique de Firth.

2.3.3. le principe d'idiome

Sinclair accordait beaucoup d'importance à la phraséologie [Sinclair, 2004 : 29] notamment parce qu'il pose le problème du statut de l'unité lexicale (cf. *infra* 2.2.3). Pour Sinclair, les mécanismes de production et d'interprétation du langage peuvent se décrire par deux principes opposés :

> « It is contended here that in order to explain the way in which meaning arises from language text, we have to advance two different principles of interpretation. » [Sinclair, 1991 : 109]

Le principe de libre-choix correspond à l'idée qu'un texte est le produit de choix en cascade (hiérarchique ou linéaire) qui agissent comme des contraintes sur le type de mots pouvant apparaître à une position donnée. D'après lui, cette vision est la plus courante en analyse linguistique. Il envisage des contraintes de différente nature qui rentrent dans ce paradigme. Par exemple, il considère que la réalité exerce dans une certaine mesure des contraintes sur l'organisation linguistique d'un texte :

> « To some extent, the nature of the world around us is reflected in the organization of language and contributes to the unrandomness. Things which occur physically together have a stronger chance of being mentionned together » [*ibid.* : 110]

Il n'en dit pas plus. On peut qualifier ces restrictions de « référentielles » ; nous en avons vu un exemple à travers le principe de métonymie intégrée proposé par G. Kleiber qui permet d'expliquer comment une partie peut être désignée par le tout (cf. *infra* 1.2.5) : la proximité spatiale d'objets, et leur organisation peuvent se traduire en contraintes dans le texte.

Le registre (type et genre de texte) limite également les choix linguistiques [*ibid.*] : le niveau de langue par exemple. Enfin, les structures syntaxiques contraignent le type de mot possible dans chacune de ses positions (la position *Déterminant* dans un syntagme nominal exclura les unités de classe *Nom* par exemple). Sinclair résume ces contraintes par le terme de grammaticalité (la conformité à la structure) :

12 Ces critères d'ordre discursif seront notamment retenus et développés par Hoey pour définir le contexte d'une unité lexicale [Hoey, 2005].

> « This is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness. » [*ibid.* : 109]

Mais, malgré ces contraintes, nous dit l'auteur, il existe encore de nombreuses manières de nous exprimer. Pourtant on observe par exemple en corpus que les mots n'apparaissent que dans des structures grammaticales limitées [*ibid.* : 112]. Par conséquent ce principe est insuffisant pour expliquer les phénomènes langagiers. Il propose le principe d'idiome :

> « there is still far too much opportunity for choice in the model, and the principle of idioms is put forward to account for the restraints that are not captured by the open-choice model. » [*ibid.* : 110]

Le principe d'idiome est invoqué pour justifier l'intérêt d'une analyse qui se focalise sur l'unité lexicale. Lorsqu'on se place dans une telle perspective, on observe des phénomènes qui sont difficilement généralisables parce qu'ils associent des mots. Ce principe englobe donc ces phénomènes de « co-sélection » : les mots sont en apparence indépendants mais ils relèvent en fait d'un même choix, d'une même unité, d'un même sens. Ils co-dépendent les uns des autres ou, comme l'aurait dit Firth, ils partagent des dépendances mutuelles [Firth, 1957]. Ces unités « préfabriquées » ne peuvent être analysées en fonction de leurs parties, leur sens n'est pas compositionnel :

> « The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. » [Sinclair, 1991 : 110]

Les formules idiomatiques, les proverbes, les termes techniques et les collocations en sont des illustrations. Leurs prédominance observée en corpus justifie, pour l'auteur, l'importance de ce principe.

Pour justifier ce phénomène, l'auteur propose une corrélation entre sens et fréquence [*ibid.* : 112–113]. Nous synthétisons les deux points importants :

- Plus un mot est fréquent, plus il est difficile de définir son sens (les mots grammaticaux comme *de*). Certains mots fréquents semblent ne pas avoir de sens dans certains contextes (comme *prendre* dans l'expression *prendre les jambes à son cou*). La fréquence d'un mot peut donc être un indice de sa dépendance sémantique, de sa « délexicalisation ». La dépendance sémantique se matérialise par la co-sélection d'unités [*ibid.* : 113].
- le sens que l'on donne intuitivement à un mot (par exemple *corps* comme *ensemble des parties matérielles constituant un organisme vivant*) n'est pas celui qui est le plus fréquemment observé en corpus (par exemple *corps* comme *groupe de personnes constitué en ensemble*). Ce sens intuitif est sémantiquement indépendant (ou n'est pas le sens délexicalisé), c'est-à-dire qu'il n'apparaît pas dans des unités préfabriquées.

Observons que le principe d'idiome se justifie principalement d'un point de vue sémantique. Il rend compte des divergences entre forme et sens : l'unité mot utilisée dans les dictionnaires n'est pas toujours pertinente pour décrire le sens parce que ces mots sont employés dans des contextes où ils contribuent à la construction d'un sens plus large, plus étendu (cf. *infra* 2.3.4). Les unités devraient être caractérisées par le sens et non pas la forme et un tel sens ne peut être observé qu'en corpus. À noter que ces unités ne sont pas totalement figées, elles montrent également une variation.

Sinclair ne s'attarde pas sur les raisons qui pourraient expliquer leur existence. Il énumère la redondance de situations extra-linguistiques (des rituels de communication), la tendance à l'économie de l'effort ou encore l'urgence de la communication en temps réel :

> « To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. » [*ibid.* : 110]

Bien qu'il ne développe pas ces aspects, on peut les lier à la notion d'unités préfabriquées (« semi-preconstructed phrases ») : dans les situations de communication, la langue est codifiée, elle est un outil qui doit être efficace pour la transmission d'information [Wray, 2002]. Par opposition, le principe du libre-choix serait plus apte à décrire le créativité linguistique [Hoey, 2005]. En tant que mécanisme d'interprétation, le locuteur alterne de l'un à l'autre [Sinclair, 1991 : 114] : il utilise plus souvent le principe d'idiome parce qu'il « reconnaît » [Firth, 1957] les opérations de co-sélection et qu'il sait les interpréter ; dans le cas contraire, le principe du libre-choix lui fournit une alternative pour les cas qui supposent un effort d'interprétation. Chez Sinclair, Interprétation et Forme sont indissociables.

La dernière contribution de l'auteur au sujet du statut de l'unité lexicale que nous souhaitons aborder est sa caractérisation des unités étendues de sens.

2.3.4. Les unités étendues de sens

Pour rendre l'image d'une unité dont le sens s'étend sur plusieurs mots, Sinclair propose le terme d'« unité étendue de sens » (désormais UES). De son point de vue, l'UES est un type d'unité lexicale beaucoup plus fréquente que le mot :

> « One hypothesis, to be explored in this chapter, is that the notion of a linguistic item can be extended, at least for English, so that units of meaning are expected to be largely phrasal. Some words would still be chosen according to the open-choice principle, but probably not very many, depending on the kind of discourse. The idea of a word carrying meaning on its own would be relegated to the margins of linguistic interest, in the enumeration of flora and fauna, for example. » [Sinclair, 2004 : 29–30]

L'adéquation mot-sens serait marginale, ce serait une exception et l'UES serait la règle. Il ne s'agit plus ici de formules idiomatiques mais d'une application du principe d'idiome. Le(s) sens d'un mot s'imprègne(nt) des contextes dans lesquels il est employé, ils sont influencés par les usages qu'en font les locuteurs :

> « Part of the supporting argument for this hypothesis is that words cannot remain perpetually independent in their patterning unless they are either very rare or specially protected (for example by being technical terms, if indeed that status offers the protection that is often claimed for it). Otherwise, they begin to retain traces of repeated events in their usage, and expectations of events such as collocation arise. This leads to greater regularity of collocation and this in turn offers a platform for specialization of meaning, for example in compounds. Beyond compounds we can see lexical phrases form, phrases which have to be taken as wholes in their contexts for their distinctive meaning to emerge, but which are prone to variation. » [*ibid.* : 30]

2.3.Lexique et sens : les travaux de Sinclair

Comme nous allons le voir, une UES n'est pas une expression figée, mais une unité lexicale abstraite qui peut avoir plusieurs formes (comme un verbe changeant de forme en se conjuguant).

Sinclair propose quatre phénomènes d'ordre sémantique qui permettent de caractériser la nature des relations qu'entretiennent les mots d'une UES :

- la collocation
- la colligation
- la préférence sémantique
- la prosodie sémantique

L'auteur se propose de décrire l'UES du nœud *naked eye*, qui apparaît 154 fois dans le corpus Bank of English (211 millions de mots). Il analyse successivement l'environnement gauche et droit de ce nœud, position par position.

- Il observe par exemple que la position N-1 (première vers la gauche) est occupée à 95% par la forme *the*, soit une collocation. L'unité *the naked eye* est figée et *the* en fait partie.
- Les formes qui occupent la position N-2 sont plus variées mais 90% sont des prépositions (with, to, etc.). C'est donc un cas de collocation d'ordre grammatical ou « colligation », définie simplement comme la co-occurrence de choix grammaticaux [*ibid.* : 32].
- Les formes apparaissant en position N-3 ne peuvent pas être caractérisées en termes de collocation ou de colligation (*verbes*, *adjectif*). En revanche, elles partagent une trait sémantique commun (ou sème), celui de *visibilité* (*seen*, *visible*, *invisible*, etc.), ce qu'il nommé la « préférence sémantique » :

  « We now consider N-3, and leave on one side the short and technical instances (reducing the total number to 134). It is immediately clear that variations on two words – *see* and *visible* – dominate the picture.
  All of these are prominent collocations, restricted to the two word classes 'verb' and 'adjective'. On this occasion colligation, being divided between the two, is not as important as another criterion, that of *semantic preference*. Whatever the word class, whatever the collocation, almost all instances with a preposition at N-2 have a word or phrase to do with visibility either at N-3 or nearby. » [*ibid.*]

- L'analyse des formes apparaissant dans les positions plus à gauche sont encore plus variées. Mais d'après l'auteur, elles partagent des propriétés d'ordre pragmatique. Elles expriment une difficulté dans 85% des cas : *small*, *faint*, *weak*, *difficult* [*ibid.* : 33]. La catégorie qu'il propose est la « prosodie sémantique ». Cette notion guide l'interprétation de l'UES globale, et correspond à la fonction qu'elle remplit en contexte :

  « It expresses something close to the 'function' of an item – it shows how the rest of the item is to be interpreted functionnally. Without it, the string of words just 'means' – it is not put to use in a viable communication. So in the example here, the attention to visibility and the strange phrase *the naked eye* are interpreted as expressions of some kind of difficulty. » [*ibid.* : 34]

  Le terme de prosodie sémantique, emprunté à B. Louw [Louw, 1993] mais employé dans un

sens différent, n'est pas sans rappeler la prosodie de Firth [Firth, 1957 : 194] : ce dernier proposait que le submorphème */-sl/* en anglais était emprunt d'une forme de dégoût. Dans ce cas-ci, les mots pris individuellement n'expriment pas cette prosodie ; elle ne se caractérise que par rapport à une UES :

> « The precise extent of the prosody, and the nature of its realization, cannot be determined in advance; and once it is identified with a phrasing it will be part of the meaning even if it has no clear expression. » [Sinclair, 2004 : 37]

Sinclair en conclut que l'initiale d'une UES est difficilement détectable, à cause de la prosodie sémantique mais que la finale est fixe :

> « The beginning of the item is very difficult to detect normally, because it is so variable; on the other hand the end is fixed and obvious. But if the analysis is correct, the whole phrase must be seen as the result of a single choice, with no doubt a number of subsidiary internal choices. » [*ibid.* : 34]

En revanche, il estime que l'UES procède initialement de la prosodie et que les choix lexicaux et grammaticaux dépendent de cette prosodie :

> « The speaker/writer selects a prosody of difficulty applied to a semantic preference of visibility. The semantic preference controls the collocational and colligational patterns, and is divided into verbs, typically *see*, and adjectives, typically *visible*. […] The final component of the item is the *core*, the almost invariable phrase *the naked eye*. » [*ibid.*]

Nous avons représenté cette UES dans la figure (2.3) en mettant en vis-à-vis les catégories (en rectangle) et exemples de Sinclair (en ellipse). La succession des choix est indiquée par des flèches.

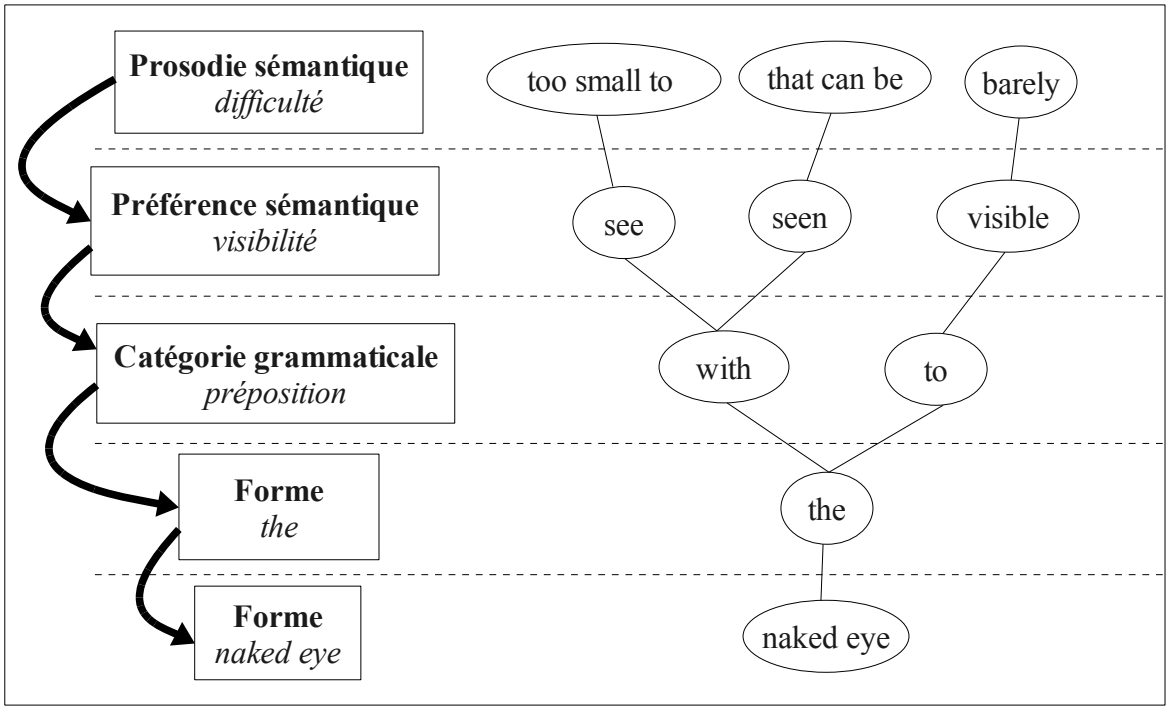


FIG 2.3 – Exemple d'Unité Étendue de Sens dont le noyau est naked eye

Sinclair indique qu'il ne s'agit que d'un type d'UES : on pourra donc imaginer d'autres successions et combinaisons de ces catégories. Résumons que les critères qu'il a employé sont la forme, la catégorie grammaticale, la catégorie sémantique et la prosodie sémantique et que son analyse s'est déroulée en étapes successives en agrandissant la taille de l'UES à chaque position vis-à-vis du nœud-cible.

Si Sinclair considère que c'est une unité lexicale, on peut également la considérer comme une grammaire qui permet d'exprimer la notion de difficulté de visibilité. En effet, en les étendant ainsi, les UES que l'on pourra créer partageront de nombreuses similarités avec les grammaires qui prennent en compte des critères sémantiques et/ou lexicaux.

2.4. Les grammaires de corpus

2.4.1. La grammaire de patron

Une des applications de cette perspective lexicale est l'éclairage qu'elle peut apporter sur la grammaire. Premièrement, les mots dits « grammaticaux », ou « mots vides », comme les déterminants, prépositions, auxiliaires, peuvent recevoir de nouvelles descriptions à partir de leurs formes de surface. La colligation, définie comme phénomène de coocurrence entre une catégorie grammaticale et une forme, ouvre la voie à la mise en valeur de structures grammaticales préférentielles pour chaque unité lexicale.

La Grammaire de Patron s'inspire largement des travaux de Sinclair et a pour ambition de départ de définir un inventaire de « patrons » à partir de l'analyse indépendante de chaque mot en corpus. Un patron est défini comme l'ensemble des mots et structures régulièrement associées au mot-cible et qui « contribuent » à son sens :

> « The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it. » [Hunston & Francis, 2000 : 37]

Les unités qui composent un patron (excluant le mot-cible) sont de deux types : formes (prépositions : *by*, *of*, … ; conjonctions : *that*, … ; etc.) et catégories grammaticales (verbe, nom, adjectif, …). Un patron consiste en une séquence ordonnée de combinaisons de ces unités, tel qu'observée en corpus, sans interprétation syntaxique superflue ; ni fonction, ni relation syntaxique ne sont employées :

> « The approach to complementation patterns described in this book is new in that it focuses on the formal components of a pattern rather than on a structural interpretation of those components. For example, the coding **V n** is preferred to 'Verb plus Object' or 'Verb plus Complement'. » [*ibid.* : 151]

Les auteurs se démarquent donc d'une analyse grammaticale standard et proposent une description de patrons de « surface » qui, malgré un jeu de catégories limité, ainsi justifié, peut séduire : la caractérisation de la disposition linéaire des unités sur le plan syntagmatique, libérée de raisonnements d'ordre syntaxique, peut mettre à jour de nouvelles propriétés et structures. Un des dangers qui guettent ce type d'approche est l'éparpillement de patrons relevant d'une même structure, ce qui pose de nombreuses questions de méthodologie. Par exemple, la construction en sujet inversé dans une question devrait-elle traitée différemment d'une construction où le sujet précède son verbe ? Une autre question, liée à celle des critères d'identité d'un patron (variations sur le plan syntagmatique), est celle du degré de granularité d'un patron (variations sur le plan paradigmatique) ou des unités qui le composent. Devra-t-on par exemple distinguer les différentes formes verbales (temps, mode, aspect) et sur quels critères les distinguera-t-on ?

Les difficultés posées par une telle entreprise font légion : la volonté de décrire des patrons d'usage se confronte à la nécessité d'abstraction de toute analyse. Les auteurs reconnaissent les limites de leur entreprise dans leur analyse de la voix passive, qu'ils considèrent comme une variante non pertinente :

> « When a verb is in the passive voice, the order of the elements is different, with the Object of the active sentence functioning as the Subject of the passive sentence. As mentionned above, a strict adherence to the surface description would perhaps involve the treatment of the passive as a separate pattern. However, for the sake of convenience and simplicity, it is considered here as a variant of the active pattern. » [*ibid.* : 60]

La Grammaire de Patrons crée un répertoire de patrons pour des noms, des adjectifs ou des verbes : le patron est indépendant du mot-cible qui en est le noyau, il a une existence propre et peut donc être associé à divers mots. La question qui se pose alors est celle du lien entre noyau et patron, où en d'autres termes, celle des contraintes qu'exerce le patron sur le type de mots qui occupent la position de noyau. Les auteurs s'aventurent ainsi sur le terrain du sens en rappelant que leurs patrons ne sont pas des unités de sens :

> « So far we have talked about 'pattern' and 'meaning' as though these were separate, as though the pattern were a framework into which words with particular meanings could be slotted. This is essentially a matter of convenience: it allows us to talk about a word 'having' a pattern and to compile a dictionary entry for a word which lists the patterns it 'has' (this is done in CCED). Moreover it allows us to generalize about patterns, and to list them as if they existed as an entity apart from the words that occur at their core. This approach, however, runs counter to the work of Sinclair, for example, whose investigation into the behaviour of particular lexical items […] stress the uniqueness of each 'meaning unit'. We would come closer to the spirit of Sinclair's work if we defined a pattern as a sequence of elements including the core. » [*ibid.* : 86]

Étant donné que les patrons ne sont pas véritablement composés de catégories sémantiques (formes, essentiellement des mots grammaticaux, et catégories grammaticales), les auteurs évaluent dans quelle mesure ils contraignent le sens du mot-noyau. Ils analysent donc la cohérence sémantique de la liste des mots-noyau (leurs formes et non leurs occurrences) considérés comme collocats du patron. L'expérience n'est pas concluante : dans la grande majorité des cas, un patron ne réunit pas un groupe de sens. Il isolent par exemple 19 groupes de sens (celui qui regroupe le plus de formes est par exemple *augmentation, réduction ou changement* avec des formes comme *acceleration*, *cut* ou *trend*) pour le patron *N in n* (*/Nom$_1$ dans Nom$_2$/* ; [*ibid.* : 90–95]), où c'est la cohérence sémantique du premier nom (en majuscule) qui est analysé. Certains groupes sont simplement intitulés *autre noms* (« other nouns » ; [*ibid.* : 93]) et parfois les noyaux n'entretiennent aucun rapport [*ibid.* : 84].

Des patrons qui ne sont pas établis sur des critères sémantiques ne parviennent pas à contraindre le sens de leurs mots-noyau : aurions-nous dû nous attendre à une conclusion différente ? Comme le rappellent les auteurs (citation plus haut), il faudrait peut-être prendre en compte la sémantique de ce noyau, mais il ne serait plus possible de comparer ces noyaux. Leur intention était d'identifier des patrons généraux communs à de nombreuses formes plutôt que des patrons propres à une forme [*ibid.* : 77–82]. L'expérience de la Grammaire de Patron, qui a le mérite d'avoir développé la notion de patron de surface, ouvre la voie à l'utilisation d'autres catégories ou indices contextuels. L'approche Corpus Pattern Analysis (CPA) en LCI fera directement appel à des catégories ontologiques, c'est-à-dire référentielles.

2.4.2. La grammaire de patrons sémantiques CPA

CPA [Hanks, 2008] est présentée comme une méthode d'analyse de corpus, une « technique », le terme désignant également le nom du projet qui met en œuvre cette méthode.

En tant que méthode, CPA permet de construire un dictionnaire de patrons de verbes de la langue anglaise, une nouvelle génération de dictionnaires qui répondrait aux objectifs suivants :

- Décrire le sens d'un verbe en fonction des patrons principaux dans lesquels il est employé.
- Ordonner les sens des patrons en fonction du critère de fréquence observée en corpus.
- Recenser uniquement les usages identifiés dans des corpus de texte.

Cette méthode s'inspire de modèles sémantiques comme le Lexique Génératif [Pustejovsky, 1998] et la Sémantique des Préférences [Wilks, 1975].

L'objectif du dictionnaire de patrons est de fournir un inventaire des principaux patrons des principaux verbes de l'anglais (« all the normal patterns for all the normal verbs in English » ; [Hanks & Ježek, 2008 : 391]) en analysant des échantillons de concordances issues de larges corpus (le BNC par exemple). Hanks se démarque de la grammaire de patron en remarquant que les patrons CPA sont des patrons de sens :

> « H&F limit themselves to expressing patterns all at more or less the same level of generalization, almost exclusively in terms of word classes (parts of speech), with the exception of certain prepositions. CPA, by contrast, devotes a great deal of attention to selecting the appropriate level of generalization to capture the meaning of the lexical pattern and to contrast it with other meanings activated by other patterns for the same verb. » [Hanks, 2008 : 111]

Dans le modèle CPA, chaque patron est associé à un sens (nommée *implicature*), c'est une unité sémantique. La conséquence logique est que la polysémie d'un verbe est fonction du nombre de ses patrons.

Pour « capturer le sens du patron », Hanks entend mieux caractériser les éléments qui le constituent. L'originalité de son approche repose sur un usage important de catégories ontologiques nommés *types sémantiques*. Le type sémantique est une catégorie générale issue d'une ontologie surfacique (en l'occurrence, une version simplifiée de la Brandeis Semantic Ontology), représentant une propriété partagée d'une classe d'entités. Par exemple, le type [[Humain]] renverra à *homme*, *femme*, *enfant*, etc., i. e. toutes les unités lexicales partageant les propriétés d'un être humain :

> « [...] we all know, informally, that the expression "John Major" falls into the class [HUMAN]. But a satisfactory formal account of such classes is not yet available. Preliminary empirical work suggests that all a-priori assumptions are suspect. For example, the class [HUMAN] seems plausible enough, but it may turn out to be unsatisfactory. As a matter of syntax, it may work better if divided into two classes: defined, on the one hand, by properties which John Major shares with cats, horses, and monkeys, such as eating, sleeping, and climbing [i.e. the type ANIMAL], and on the other hand by properties which he shares with nations, governments, business organizations, family-history societies, and computers, i.e. the type [COGNITIVE], namely analysing, negotiating, banking money, making statements, expressing sympathy, and so forth. The details are uncertain, so for present purposes it will suffice to use [HUMAN] or [PERSON]. » [Hanks, 1997 : 8]

La liste de types sémantiques dont il dispose en comprend 65 et ceux-ci sont sélectionnés en fonction de leur « prévalence » dans les patrons :

> « The Brandeis Shallow Ontology (BSO) is a shallow hierarchy of types selected for their prevalence in manually identified selection context patterns. At the time of writing, there are just 65 types, in terms of which patterns for the first one hundred verbs have been analyzed. New types are added occasionally, but only when all possibilities of using existing types prove inadequate. Once the set of manually extracted patterns is sufficient, the type system will be re-populated and become pattern-driven. » [Pustejovsky et al., 2004 : 927]

Les catégories ontologiques sont structurées et hiérarchisées en fonction de la nature des entités, des événements ou des propriétés et chacune de ces catégories majeures est (à peu de choses près) le corrélat sémantique des catégories de nom, de verbe, ou d'adjectif, respectivement. En conséquence, elles peuvent être considérées comme des sous-catégories de ces catégories grammaticales majeures. La liste des types est reproduite en figure (2.4).

FIG 2.4 – Ontologie surfacique BSO [*ibid.*]

Le type sémantique n'est pas la seule catégorie mise à profit dans les patrons sémantiques. Hanks cherche également à décrire les indices contextuels qui exercent une influence sur le sens du patron.

Il peut alors s'appuyer sur [Pustejovsky et al., 2004] :

- Les catégories comme [Adverbe] ou des syntagmes grammaticaux comme [Clause].
- Des catégories fonctionnelles (peu décrites), comme des sous-types de proposition (proposition de but).
- Les classes d'unités nommées « catégories sous-valencielles », comme la présence/absence de déterminants (comparer *take place* et *take his place*).
- Les rôles sémantiques, qui sont des informations sur le rôle (*judge*, *driver*, par exemple) que remplit un des éléments du patron en contexte, rôle qui n'est pas hérité par son type sémantique.
- Si aucune de ces étiquettes ne caractérise pertinemment l'élément, le lemme ou la forme de l'unité est employée, ce qui peut se produire pour des expressions idiomatiques figées fréquentes.

L'exemple (22) illustre l'usage de certains de ces indices pour un patron du verbe *to scratch* [Hanks, 2008 : 121], qui signifie (implicature) qu'un être humain essaie à bénéficier de quelque chose dans des circonstances difficles.

(22)[[Human]] scratch [NO OBJ] {around | about} {for [[Entity = Benefit]]}
IMPLICATURE : [[Human]] tries to obtain [[Entity = Benefit]] in difficult circumstances

Dans cet exemple, *Human* et *Entity* sont des types sémantiques, *Benefit* est le rôle sémantique, la catégorie NO OBJ indique que le verbe ne se combine pas avec des objets directs et *around*, *about* et *for* sont des formes. Pour donner un exemple d'usage de catégorie grammaticale, on peut par exemple s'intéresser au verbe *to urge*,(dans le sens de *presser quelqu'un à faire quelque chose*) qui se combine avec un syntagme prépositionnel dont la tête syntaxique est un verbe (exemple 23 extrait de [Hanks et al., 2007 : 6]).

(23)[Human 1]] urge [[Human 2]] {to [V]}

Cela signifie que l'auteur n'a pas trouvé de propriété sémantique plus pertinente que le verbe.

Tous ces éléments peuvent jouer un rôle dans la formalisation du patron (la ressource qui en est le résultats contient 718 verbes au 14 septembre 2011[13]). Les patrons principaux de chaque verbe peuvent correspondre à de nombreuses réalisations (variations) en contexte mais conservent le même sens. En d'autres termes, le lexicographe doit écarter les variations qui n'ont pas d'influence sur le sens. Ainsi, Hanks ne fait figurer la variation induite par la voix passive que lorsque celle-ci contribue à la caractérisation d'un sens particulier du patron, ce qui constitue une avancée par rapport à la Grammaire de Patron. Mais ces patrons ne peuvent plus prétendre être des patrons de surface ; ils sont d'ailleurs très proches des structures prédicatives.

L'usage des catégories ontologiques fait naître des problèmes spécifiques à cette méthode : le problème majeur est celui de la métonymie auquel sont sujets les arguments du patron (cf. *infra* 1.2.5). Le choix du type sémantique (parmi un ensemble de possibles) s'établit à partir de l'analyse des collocats d'un verbe dans une position donnée. Hanks utilise une interface d'interrogation de corpus, le SketchEngine, qui les génère automatiquement. Au vu de la variété des unités qui peut surgir dans cette liste, le lexicographe établit le choix qui lui semble le plus approprié. Si plusieurs

13 http://deb.fi.muni.cz/pdev/

groupes semblent se démarquer, ce qui est appelé une alternance de type (« alternation of semantic type » ;[Hanks & Ježek, 2008 : 397]), il est possible (1) de différencier les patrons en fonction de cette alternance (2) ou de lister les types sémantiques pertinents dans le patron.

1. Un patron est généralement créé lorsqu'un glissement sémantique récurrent peut être identifié, c'est-à-dire lorsque la variation de type semble corréler avec une variation de sens du patron. Deux patrons peuvent ne varier par exemple qu'en fonction du type sémantique en position *objet*. Les exemples que compare Hanks sont *to fire a person* (*virer quelqu'un*) et *to fire a bullet* (*tirer une balle*), dans lesquels la variation du type ([[Human]] contre [[Projectile]]) corrèle avec la variation de sens du patron. *Projectile* ne fait pas partie de l'ontologie figure 2.4, mais Hanks affirme que cette ontologie est ajustée progressivement à partir de ses analyses de corpus.
2. Il est également possible qu'une alternance de type soit observée mais n'entraîne pas de variation sémantique du patron. Pour rendre compte de cette alternance de type, il faut s'assurer qu'il n'existe pas un type sémantique plus abstrait mais plus approprié qui convienne pour décrire la totalité des mots observées dans une position donnée. Si tel n'est pas le cas, Hanks propose tout de même d'inscrire cette alternance dans le patron en spécifiant dans la position concernée, la liste des types sémantiques possibles. Le type le plus fréquemment rencontré apparaîtra en premier et l'autre (ou les autres) constituera un type alternatif également autorisé dans le patron. Les alternances de type ne provoquant pas la création d'un patron peuvent s'expliquer en contexte par des phénomènes métonymiques, que CPA ne traite pas. La métonymie correspond au phénomène dans lequel une chose est désignée par la dénomination d'une autre avec laquelle elle entretient une relation sémantique, comme l'individu par l'organisation dont il fait partie (*Matignon* pour *premier ministre*) ou le contenu par son contenant (*verre* pour *eau*). Ces relations ne donnent pas lieu à la création d'un nouveau patron :

> « Alternations are linked to the main CPA pattern through the same sense- modifying mechanisms as those that allow for exploitations (coercions) of the norms of usage to be understood […] Note that alternations are different realizations of the same norm, not exploitations (i.e., not coercions). » [Pustejovsky et al., 2004 : 926]

À travers cette citation, Hanks rappelle que c'est le sens du patron qui prime sur les variations de sens de ses éléments et il rejoint Sinclair sur ce point. La différence se situe néanmoins sur l'origine des catégories : alors que Hanks manipule des catégories de sens compartimentées par catégorie grammaticale, les catégories sémantiques de Sinclair sont plus proches des traits communs et de notions (« visibilité » par exemple). Or, il est tout à fait possible qu'une notion sémantique qui apparaît fréquemment dans un patron, comme la notion de force (*strength*) liée à un mot comme *argue* soit exprimée par des catégories grammaticales différentes (*strong*, *strength*, *strongly*, …).

Les patrons CPA relèvent donc d'une représentation syntaxico-sémantique et non uniquement sémantique. Un aspect non négligable de ce modèle est qu'il cherche à identifier des normes de la langue plutôt qu'à lier ces patrons à des pratiques socio-discursives ou à des genres de textes particuliers. On retrouve ici une préférence pour le système de la langue de de Saussure (cf. *infra* 1.3.1) : l'usage du corpus, malgré le soucis de détail apporté à la description linguistique, n'est qu'un intermédiaire, un moyen pour atteindre un objectif de généralité. Cette hypothèse est justifiée par l'usage de corpus généraux de référence (comme le British National Corpus ou BNC), qui mêlent des textes de type et de genre différents : en procédant ainsi, on s'abstrait des spécificités des textes

pour identifier ce qui est commun aux occurrences d'un mot-cible. La taille du corpus est importante mais insuffisante puisque l'on pourrait concevoir un large corpus d'un genre unique qui nous permettrait d'identifier ses normes. C'est bien la variété des textes qui peut justifier l'identification de normes générales d'une langue. À ce titre, les linguistes de corpus posent tous la même question : qu'est-ce qu'un corpus représentatif ou équilibré et comment le construire (cf. *infra* 2.1.3) ? Pour ce qui concerne CPA, le problème est plus particulièrement de savoir s'il n'y a pas un conflit entre l'identification de normes générales et la création de patrons sémantiques : Peut-on véritablement identifier des normes en intégrant des propriétés sémantiques dans les patrons ? Une initiative doit être relevée de ce point de vue, celle des grammaires de fonction.

2.4.3. Les grammaires de fonction

Ce que nous appelons les grammaires de fonction renvoie à des travaux inspirés des analyses de Sinclair sur les cadres collocationnels. Les cadres collocationnels (« collocational frameworks » ; [Renouf & Sinclair, 1991]) désignent des patrons contrôlés par des mots grammaticaux (*un*, *de*, etc.). Il s'agit plus exactement d'analyser la nature des mots apparaissant entre deux formes grammaticales, comme */an + ? + of/*. Ces cadres ont été discutés plus en détail (sans les nommer) dans l'analyse que Sinclair fait de *of* [Sinclair, 1991 : 83], qui est une des formes les plus fréquemment employées en corpus. La difficulté de l'analyse d'une forme comme *of* réside en la quantité de données à analyser et l'auteur propose une méthode pour la contourner [*ibid.* : 84]. *of* apparaît pour 80% des cas dans des syntagmes nominaux de la forme */N of N/*. L'auteur constate que pour la majorité, la tête sémantique de ce groupe n'est pas le premier nom, comme on pourrait le croire. La fonction de *of* est d'introduire un second nom, qui, dans la majorité des cas, est le plus « saillant » :

> « To begin with, we note that in most cases the second noun (N2) appears to be the most salient. » [*ibid.* : 85]

En effet, l'omission du second nom est responsable d'une perte de sens plus critique alors que le premier sert plus souvent de spécifieur ou de focalisateur, c'est-à-dire qu'il peut être omis sans que le sens principal soit altéré (par exemple le nom *millions* dans le groupe [*millions of cats*]). La sémantique de ce syntagme repose par conséquent non sur la nature (valeur numérique, valeur temporelle, partie du corps, etc.) partagée par les noms en relation mais sur la fonction qu'exerce l'un sur l'autre. Un des patrons qu'il identifie est */[Focus noun] of [Focused noun]/* et il propose de préciser la fonction que remplit le premier nom [*ibid.* : 96–97] :

- Nom de focalisation (« Focus nouns ») : *N1 permet de focaliser sur un aspect de N2*
  - N1 désigne une mesure conventionnelle (quantifieurs) : *millions of cats*
  - N1 désigne une mesure moins conventionnelle (partitifs) : *the bottle of port*
  - Focalisation sur une partie : *the end of the day*
  - Focalisation sur une partie spécifique : *the first week of the war*
  - Focalisation sur un composant, aspect ou attribut : *the text of two or three White House tapes*

Ces fonctions qui sont ici appliquées au syntagme nominal sont plus fréquemment employées dans le domaine du verbe, avec les fonctions syntaxiques (*sujet*, *objet*, *complément de lieu*, etc.) ou

encore les rôles sémantiques (*Agent*, *Patient*, *Bénéficiaire*, etc. ; cf. *supra* 4.3.1). Ajoutons que l'étude de Sinclair est limitée à une relation binaire de deux noms liés par la forme *of*, alors que l'analyse des verbes demande souvent la prise en compte de relations ternaires ou plus (comme pour *quelqu'un demande quelque chose à quelqu'un d'autre*). Ce type de description sémantique a été approfondi dans deux directions : les grammaires de définition et les grammaires d'évaluation.

L'objectif de la grammaire de définition [*ibid.* : 123–137] ; [Barnbrook, 2002] était de décrire les formes que prennent les définitions dans le dictionnaire. Une analyse fonctionnelle est de ce point de vue intéressante puisque tout type de mot peut apparaître : on ne peut par conséquent s'appuyer sur la nature sémantique des éléments de la définition comme contraintes. Les définitions d'un dictionnaire sont codifiées et leur analyse permet de dégager des régularités (formules phraséologiques ou police typographique) :

> « A dictionnary is a text in which most units of discourse are very brief, and in which the overall structure is highly repetitive. The user will naturally be led to expect that the same kind of type-face, or the same kind of phraseology, will carry the same sort of information. » [Sinclair, 1991 : 134–135]

C'est en contraignant ici le type de texte que Sinclair propose de décrire la forme des définitions. Une définition peut se formuler de différente manières dans un dictionnaire, mais on peut isoler deux parties, celle qui présente l'unité définie (le « topic ») et celle qui l'explique (le « comment ») : *topic* et *comment* sont leurs fonctions. Ces parties sont en fait un peu plus complexes car des opérateurs sont insérés pour lier ces deux informations, tel qu'illustré figure (2.5).

FIRST PART				SECOND PART		
OPERATOR	CO-TEXT (1)	TOPIC	CO-TEXT (2)	OPERATOR	COMMENT	CHUNKS
	a	house		is	a building in which people live	1 2
if	you	defeat	someone		you win a victory over them in a contest such as …	1 2
	a	pure	substance	is	not mixed with anything else	
if	something happens	often			it happens many times or much of the time	1 2

FIG 2.5 – Exemple de forme de définition [ibid. : 125]

G. Barnbrook (Barnbrook, 2002) a utilisé ce modèle pour implémenter un système qui puisse segmenter automatiquement les entrées d'un dictionnaire COBUILD (le *Collins Cobuild Students' Dictionary*). Son système permet de reconnaître six types de définitions (figure 2.7) et permet ainsi de traiter 53% des entrées de ce dictionnaire (figure 2.6).

Type	Number
1	9404
2	580
3	4249
4	1826
5	161
6	575
Total 16,795 or 53.5% of the total.	

FIG 2.6 – Nombre de définitions reconnues [Barnbrook, 2002]

Comme on l'observe dans la figure 2.7, Barnbrook a détaillé la nature fonctionnelle des éléments de la définition (à gauche sur la figure) : une définition met en relation un mot (« headword ») et sa définition (« explanation ») dans le cadre d'une structure identifiée par des mots-clés (« hinge », « co-text »). Il ajoute par exemple pour le premier type (*a churchyard is an area of land around a church where dead people are buried*) la catégorie *super-ordonné*. Un *super-ordonné* (« superordinate ») peut apparaître dans une définition pour désigner l'hypéronyme du *headword*. Lorsque ce super-ordonné apparaît, il est suivi d'un discriminateur (l'« accidens » au sens aristotélicien) qui le distingue de ses co-hyponymes.

```
Operator : a
Headword : churchyard                                   1
Hinge    : is
Match    : an
Superordinate : area
Discriminator : of land around a church where dead
people are buried.
```

```
Operator : if
Cotext   : you are                                      2
Headword : flabbergasted
Match    : you are
Explanation : extremely surprised;
```

```
Operator : if
Cotext   : you                                          3
Headword : manhandle
Cotext2  : someone
Match    : you
Explanation : treat *them* very roughly
```

```
Cotext   : something
Operator : that                                         4
Hinge 1  : is
Headword : plush
Hinge 2  : is
Explanation : smart, comfortable, and expensive.
```

```
Operator : if
Cotext   : you do something                             5
Headword : thankfully
Match    : you do it
Explanation : feeling happy and relieved that something
is the case or that something has happened.
```

```
Headword :vastly                                        6
Hinge    : means
Explanation :very much or to a very large extent.
```

FIG 2.7 – Les six types de définition [ibid.]

Le principal intérêt de cette grammaire est d'associer des informations sans avoir à définir par avance la nature des unités mises en relation. L'identification des catégories suppose uniquement une segmentation en mot et la définition d'un lexique d'indices contextuels correspondant à des mots (*if*, *means*, *something*) ou des séquences de mots (*you do something*, *you do it*). Il serait intéressant de projeter cette grammaire sur d'autres textes que des dictionnaires : après tout, les dictionnaires ne sont pas les seuls types de texte dans lesquels on peut trouver des définitions. La définition devient alors un type d'énoncé [Rebeyrolle, 2000] uniquement caractérisé par sa fonction discursive, ce que nous appellerons une notion.

S. Hunston et J. Sinclair expérimentent la construction d'un tel type de grammaire fonctionnelle sur la notion d'évaluation. Comment exprime-t-on nos jugements sur les choses et quels types de jugement exprime-t-on ? L'évaluation couvre un grand nombre de concepts comme l'affect (les émotions), le jugement (l'évaluation morale) ou encore l'appréciation (la qualité esthétique) [Martin, 2000], mais aussi la subjectivité, la modalité ou encore l'opinion. Elle est, comme la définition, difficile à caractériser parce qu'elle se réalise par une variété de formes ou de catégories grammaticales :

> « Evaluative language presents difficulties in analysis because there is no set of language forms, either grammatical or lexical, that encompass the range of expressions of evaluation. » [Hunston, 2011 : 3]

Un grand nombre d'unités lexicales expriment cette notion et il est possible de discriminer des patrons communs en fonction de la catégorie grammatical de la catégorie évaluative (qui porte la nature de l'évaluation). Par exemple la catégorie évaluative de l'exemple (24) est *frustrating*.

(24)It’s frustrating when people try to do things and are held up with red tape.

La structure dans laquelle est employé cet adjectif est de type */it 's ADJ when CLAUSE/* et peut être utilisée pour de nombreux autres adjectifs, comme *better*, *great*, *funny*, etc.

Les fonctions principales de la grammaire d'évaluation sont l'*entité évaluée* et la *catégorie évaluative*, auxquels peuvent s'ajouter l'*évaluateur* ou des *restrictions sur l'évaluation*. Hunston et Sinclair [Hunston & Sinclair, 2000 : 99] distinguent 33 structures dont deux sont illustrés dans la figure (2.8).

Evaluation carrier		Evaluative category	Thing evaluated
it	v-link	n	to-inf
It	was	a damn nuisance	to have to put on new clothes and go out.

Thing evaluated		Evaluative category	Person affected
	v-link	N	*for* n
They	turned out to be	a nuisance	for match anglers.

FIG 2.8 – Exemples de structures d'évaluation [Hunston, 2003 : 348]

On remarque, sur la figure (2.8), que ces structures correspondent aux patrons grammaticaux que nous avons décrit en abordant la Grammaire de Patron (cf. *infra* 2.4.1) dont l'une des positions est une catégorie évaluative. Aucun autre indice ne peut être employé dans ces cas.
Les auteurs envisagent la possibilité d'implémentation d'une telle grammaire, mais se focalisent sur sa formalisation, *i.e.* le travail en amont. L'évaluation n'est qu'une illustration du potentiel de ces grammaires, qui constituent le lieu où se rencontrent grammaire et lexique comme ils le soulignent :

> « This study of evaluation offers some support for the assertion that large quantities of text would be amenable to analysis using local grammars, and that such an analysis would be more simple, more precise, and more useful than an analysis using a general grammar. It would be simple in that each local grammar might need to be fairly extensive. It would be precise in that each local grammar could be stated in its own terms, without the need to fit with more general statements. It would be useful because the terminology used would be reasonably transparent and would immediately relate the grammar and lexis of each part of the text to its discourse function. » [Hunston & Sinclair, 2000 : 101]

Quels seraient les notions possibles que l'on pourrait envisager pour une grammaire locale ? Le lecteur averti n'aura pas manqué de reconnaître des similarités entre ces grammaires locales et FrameNet (cf. *supra* 4.3.1). Une approche hybride pourrait mettre à profit l'avantage de ces deux modèles comme le propose Hunston [Hunston, 2003] : nous reviendrons sur ce point en 4.3.

Nous parvenons au terme de cet état de l'art sur la sémantique linguistique. Nous avons voulu cette partie à la fois théorique et pratique. Les travaux en linguistique de corpus complètent en effet les paradigmes théoriques présentés au premier chapitre. Par exemple, Hanks décrit les patrons de verbe de l'anglais en s'appuyant sur des catégories ontologiques qui sont référentielles (elles désignent des propriétés des objets du monde réel). La critique de la vision nomenclaturiste mot-objet chez Rastier (seul un syntagme réfère) est mis en perspective par les travaux de Sinclair sur la phraséologie. Enfin, la nécessité d'analyser les mots dans leur contexte est partagée par la sémantique textuelle et la linguistique de corpus. Cet état de l'art va nous permettre de mieux appréhender les phénomènes auxquels nous serons confrontés dans nos analyses de corpus tout au long de cette thèse.

2.5. Bilan

Comme indiqué en début de chapitre 2, la linguistique de corpus ne peut se résumer à la seule approche contextualiste inspirée des travaux de Sinclair. Nous savons que la constitution des corpus existe depuis plusieurs siècles, mais ce n'est que très récemment, notamment grâce à la numérisation des corpus, que des études systématiques ont pu être entreprises. La linguistique de corpus s'est aussi orientée vers l'analyse de discours [Stubbs, 1996] à travers l'étude de mots-clés, spécifiques à un type de discours. On retrouve une perspective similaire en France dans le cadre des travaux sur la lexicométrie [Muller, 1969], [Lebart & Salem, 1994], qui analysent la distribution statistique des mots en corpus (leurs répartitions, spécificités, etc.).

Dans ces dernières approches, l'analyse du contexte de chaque occurrence d'un mot dans un corpus est peu étudiée : les unités comptabilisées sont des mots (formes-types) et ce sont essentiellement des listes de mots qui sont analysées. L'analyse de cooccurrences, lorsqu'elle est entreprise, s'établit à partir d'une interprétation préalable des mots : la question du rapport entre les mots et leurs sens en contexte est peu discutée et le contrôle de l'interprétation n'est que rarement linguistiquement motivé. C'est pour cette raison et parce que nous jugeons qu'une linguistique de corpus doit en premier lieu s'intéresser au contexte de chaque instance d'un mot que nous nous sommes intéressé aux travaux de Sinclair.

Cette présentation nous a permis de développer la riche problématique de la définition d'une unité lexicale : la prise en compte du contexte ouvre la possibilité de définir des unités d'analyse supérieures au mot, les patrons. Ces patrons reposent sur une analyse sémantique des usages récurrents de mots, qui peut être réalisée à partir d'une analyse quantitative des co-occurrences et des collocations. Selon les modèles linguistiques adoptés, nous avons vu que les patrons peuvent être décrits sur le plan grammatical (grammaire de patrons), ontologique (CPA) ou encore sémantico-fonctionnel (grammaires de fonction).

Dans le cadre de nos recherches, nous avons choisi de nous intéresser en détail au modèle ontologique parce qu'il pose un lien explicite entre Texte et Référence, tel que nous l'avions proposé en 1.3 pour réconcilier les deux traditions linguistiques majeures. L'hypothèse sous-jacente au modèle CPA est que les propriétés référentielles des objets désignés par des expressions linguistiques se traduisent par des contraintes linguistiques de composition des mots d'un texte. Une seconde raison est qu'une des tâches du domaine d'Extraction d'Information que nous aborderons chapitre 5 vise l'extraction d'expressions linguistiques désignant des catégories sémantiques d'ordre ontologique. Il nous a donc paru nécessaire de développer une analyse linguistique de la notion de référence en chapitre 1 pour permettre de situer cette tâche dans le cadre d'une approche linguistique.

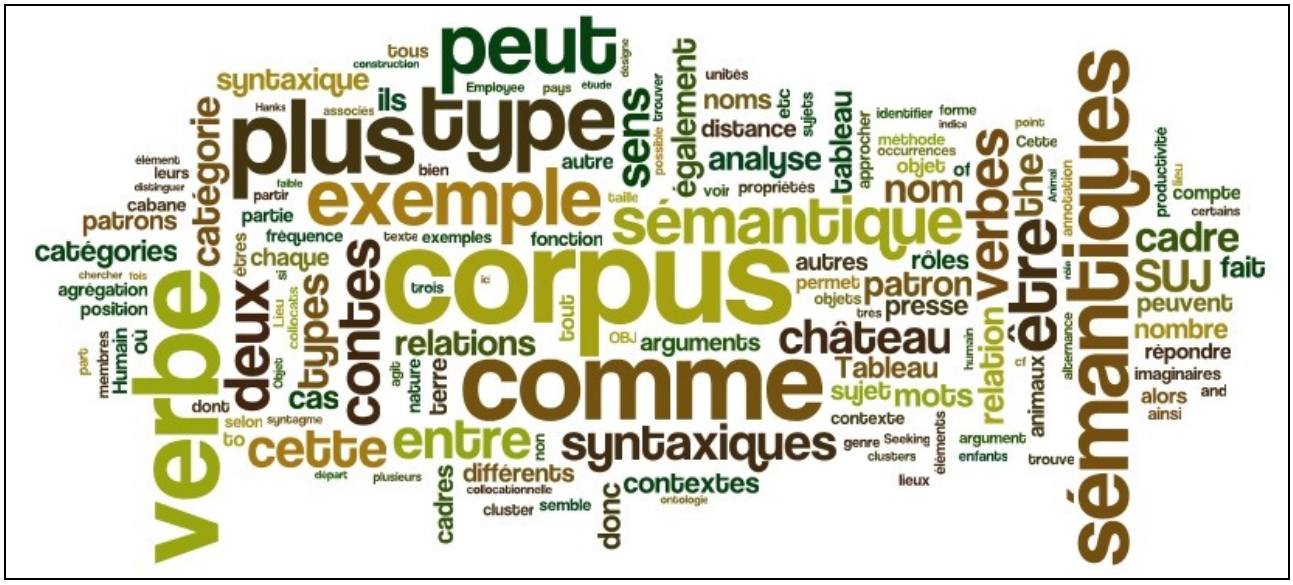
Au travers de ce volet, nous avons abordé de nombreux concepts qui auraient mérités d'être mis en perspective avec d'autres approches. Par exemple, la notion de collocation a été largement abordée en linguistique et de nombreux critères et typologies existent pour caractériser linguistiquement ces expressions figées ou semi-figées (voir par exemple [Dubreil, 2008]). Ayant présenté le principe d'idiome, nous considérons avec Sinclair que la collocation participe plus généralement d'une dimension phraséologique du langage. Les modèles sémantiques de référence de cette thèse sont la Sémantique des Cadres et le Lexique Génératif (abordé via le modèle CPA) parce que nous nous concentrons sur deux catégories majeures que sont les types et les rôles sémantiques. De nombreux autres modèles apportent des outils utiles à la description sémantique comme les primitives conceptuelles de la Sémantique Conceptuelle de R. Jackendoff [Jackendoff, 1985] ou encore les fonctions lexicales de la théorie Sens-Texte de I. Mel'cuk [Mel'cuk, 1998]. Plus généralement, nous avons choisi de limiter notre présentation à une analyse sémantique de la référence en excluant d'autres fonctions langagières comme les dimensions phatique, ou encore

conative (voir par exemple [Jakobson, 1963]). Ceci s'explique une fois de plus par notre volonté d'explorer les liens entre Texte et Référence ainsi que par l'importance de la référence dans le domaine de l'Extraction d'Information.

SECOND VOLET

SÉMANTIQUE ET TEXTE : L'ÉPREUVE DES CONTES

Les corpus ont révolutionné les pratiques d'analyse linguistique en donnant naissance à des Linguistiques de corpus, qui se définissent principalement par leur recours à des données attestées. Nous nous interrogeons dans ce volet sur la nature de l'éclairage que peut apporter une linguistique de corpus à la sémantique lexicale. Nous nous intéresserons plus particulièrement aux points de rencontre entre la sémantique lexicale et la sémantique textuelle, entre Lexique et Discours. La question à laquelle nous tâcherons de répondre est la suivante : quels problèmes peut-on rencontrer lorsque l'on cherche à appliquer un modèle de sémantique lexicale aux textes ? Cette partie se focalisera sur les relations sémantiques entre Verbe et Nom dans un corpus de contes pour enfants.



3. Analyse collocationnelle

Nous avons décrit chapitre 2 de nombreuses techniques de description lexicale en linguistique de corpus. Parmi celles-ci, l'analyse collocationnelle présente un intérêt particulier parce qu'elle fait émerger des associations significatives entre mots. Nous proposons dans ce chapitre d'appliquer cette technique pour décrire les relations entre verbe et nom sur un corpus de contes (présenté en 3.1.2) dans l'objectif de caractériser les connaissances nécessaires à un système symbolique de détection des émotions (présenté en 3.1). Nous présenterons tout d'abord les difficultés auxquelles nous avons été confronté pour la construction d'un réseau de concepts (3.2), en soulignant les intérêts et les limites de cette approche. Nous proposerons dans un second temps d'utiliser les relations syntaxiques pour compléter cette analyse (3.3) et nous évaluerons les résultats obtenus avec des ressources existantes pour le français. Ces travaux ont été réalisés dans le cadre du projet Emotirob décrit en première partie.

3.1. Contexte de recherche

3.1.1. Le Projet Emotirob

Le travail présenté dans cette partie a été établi en marge du projet EmotiRob[14], collaboration inter-disciplinaire entre les laboratoires Valoria, LI et HCTI-Adicore. L'objectif du projet EmotiRob [Saint-Aime et al., 2007] était de concevoir un robot compagnon autonome susceptible d'apporter du réconfort à des enfants fragilisés (notamment en cas d'hospitalisation longue). Le projet s'articulait en deux parties : la conception d'un module de détection des émotions des propos de l'enfant et celle d'un module d'interaction émotionnelle chargé de générer une réponse émotionnelle appropriée par des mouvements du corps, des traits du visage et des sons. Six émotions primaires devaient pouvoir être simulées : joie, tristesse, dégoût, peur, surprise et colère.

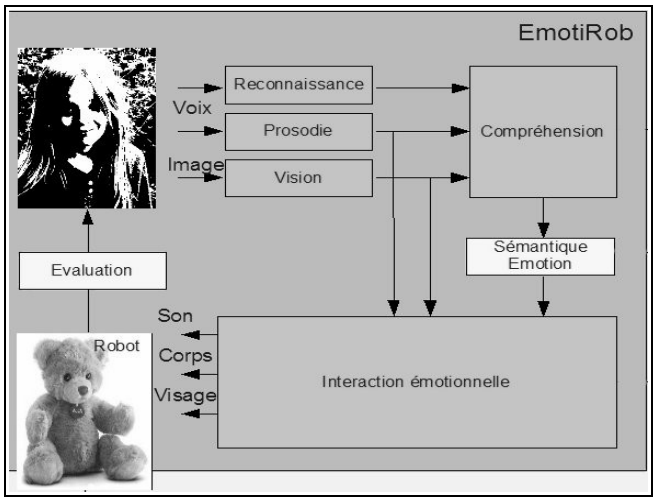


FIG 3.1 – Synoptique du projet EmotiRob

Emologus [Le Tallec et al., 2010], le module de détection des émotions élaboré dans Emotirob, s'appuie sur une détection linguistique des émotions. L'émotion véhiculée dans un tour de parole est calculée à partir d'une représentation sémantique des mots reconnus de l'énoncé. Notre recherche s'est focalisée sur la description des connaissances nécessaires à cette phase de compréhension et à l'adaptation du système de compréhension Logus [Villaneau, 2003], initialement conçu pour un tout autre domaine d'application. En d'autres termes, nous avons consacré nos recherches au vocabulaire susceptible d'être employé dans ces échanges et à son organisation sémantique.

Logus constitue une mise en œuvre d'une approche logique dans un contexte de dialogue. Son architecture repose sur la définition d'une « connaissance sémantique » du domaine d'étude, qui comprend une ontologie ainsi que les liens sémantiques entre les concepts de cette ontologie. Logus ne disposant d'aucune connaissance sur les concepts susceptibles d'être utilisés dans un tel contexte d'interaction, nous avons cherché à les identifier et à les caractériser à partir d'une étude de corpus.

14 Projet ANR (2007-2009) http://www-valoria.univ-ubs.fr/emotirob/

3.1.2. Corpus de contes

Le projet EmotiRob a été doté de deux corpus de texte. Le corpus Brassens[15] contient des retranscriptions de prises de parole d'enfants dans un contexte scolaire. La tâche remplie par ces enfants est de faire la lecture, à voix haute, d'un conte préalablement inventé en classe. Ce corpus qui comporte environ 40 000 mots, a été constitué pour tester la reconnaissance vocale. Des corpus oraux de ce type sont rares et posent des problèmes pour une analyse quantitative.

Nos recherches se sont portées sur un corpus plus large, dont la constitution posait moins de difficulté. Il s'agit d'une collection de contes pour enfants, extraite automatiquement sur la toile, comportant environ 160 000 mots. Le corpus comprend 138 contes de longueur variable (entre 120 et 17 000 mots). Les contes sont de nature diverse : la part est faite aux contes de fées classiques, comme aux contes d'animaux. Il contient également des histoires communes relatant le quotidien d'un enfant se rendant à l'école ; d'autres histoires sont situées dans un contexte historiquement daté ou encore, décrivent l'arrivée d'extraterrestres sur terre.

Les types d'auteurs varient également : le site web à partir duquel les contes ont été collectés apportait, pour la majorité, des informations quant au contexte de la création du conte. Par exemple, certains contes sont regroupés autour d'une thématique étudiée en classe et il était demandé aux enfants de broder une histoire autour de quelques éléments (à partir de personnages-types, par exemple, comme le tigre). Cette dernière tâche était similaire à celle qui a été définie dans le corpus oral. En collectant ces informations, nous avons pu nous forger une idée globale des types d'auteur de ces contes (tableau 3.1).

Type d'auteur	Fréquence	Proportion
Conteur moderne adulte	63 217	39%
Enfant	53 109	34%
Inconnu	34 314	21%
Conteur classique	9 900	6%
Total	160 540	

Type d'auteur	nb de contes	Proportion
Enfant	70	51%
Inconnu	37	27%
Conteur moderne adulte	24	17%
Conteur clasique	7	5%
Total	138	

Tableau 3.1 – Auteurs dans le corpus de contes en termes de fréquence et de nombre de textes

On observe que bien que le corpus comporte davantage d'histoires écrites par des enfants que par des adultes, les proportions en nombre de mots sont relativement similaires (39% et 34%) ; ceci s'explique par le fait que les enfants écrivent généralement des histoires courtes. Certains contes sont des adaptations de conte classique comme *Le Petit Chaperon Rouge* ou *Peau d'Âne*. Il n'a pas été possible d'identifier la nature de la source pour un quart des textes (environ 20% des mots du corpus ; étiquette *Inconnu*). Tous les textes sont écrits en français contemporain et il est rare qu'une même personne soit l'auteur de plusieurs contes.

Le corpus est donc relativement diversifié en termes de contenu ainsi qu'en termes de sources et cette variation interne peut être interprétée comme un gage d'une certaine représentativité. En effet, en variant les sources, les phénomènes idiosyncrasiques propres à chaque auteur auront tendance à s'estomper pour ne collecter que des informations spécifiques au genre de texte. Ce qui contribue également à l'unité de ce corpus, c'est le fait que les contes partagent tous un destinataire commun, les enfants : ce sont en majorité de courtes histoires destinées à être lues à des enfants afin de stimuler leur imagination.

15 http://www.info.univ-tours.fr/~antoine/parole_publique/Brassens/

3.2. L'analyse collocationnelle traditionnelle

Pour étudier le lexique d'un tel corpus et faire émerger les relations sémantiques majeures entre concepts, l'analyse collocationnelle constitue une première approche (3.2.1). Les listes de verbes ou de noms n'offrent qu'une vue partielle de la nature sémantique des unités lexicales : il faut observer leur usage en contexte. Dans un second temps, l'analyse de concordances, bien qu'elle s'avère fastidieuse, permet de vérifier les hypothèses sémantiques établies lors de la première étape.

3.2.1. De la collocation à la catégorisation

Ce que nous nommons « analyse collocationnelle traditionnelle » désigne une technique couramment employée en linguistique de corpus qui consiste à s'appuyer sur l'environnement textuel d'une unité (le co-texte) pour caractériser le sens d'un mot. Plus précisément, il faut :

- définir une fenêtre de taille arbitraire autour d'un mot-cible (par exemple 5 mots à gauche et 5 mots à droite),
- extraire les collocats (mots en cooccurrence),
- et ordonner ces collocats en fonction d'un indice de pertinence, comme le Z-score ou l'Information Mutuelle.

En pratique, seuls les mots pleins sont retenus, les mots grammaticaux (déterminants, prépositions, conjonctions, etc.) et la ponctuation étant considérés comme peu informatifs. Les collocations obtenues (associations mot-collocats) sont ensuite évaluées du point de vue de leur statut sémantique (voir par exemple [Church & Hanks, 1990]). Les collocations sont calculées par la majorité des concordanciers (Xaira[16], Wordsmith Tools[17], pour les plus connus), à supposer que le corpus soit soumis dans un format particulier.

Notre objectif n'est pas tant d'évaluer cette méthode que d'analyser dans quelle mesure les collocations traduisent des relations sémantiques. Nous montrerons comment on peut regrouper les collocats en catégories sémantiques et comment ces catégories permettent de distinguer les sens du verbe. Nous nous focalisons sur les relations verbo-nominales dans le corpus de contes. Pour filtrer les noms et les verbes de la liste des collocats, nous avons utilisé les catégories morphosyntaxiques fournies par le Tree-tagger[18], un analyseur morphosyntaxique automatique basé sur des arbres de décision. La taille de la fenêtre choisie est de 5 mots à gauche et à droite, d'après la segmentation en mots réalisée par le Tree-tagger. D'après Church & Hanks, une large fenêtre mettra plus facilement en valeur des propriétés sémantiques :

> « The window size parameter allows us to look at different scales. Smaller window sizes will identify fixed expressions (idioms such as bread and butter) and other relations that hold over short ranges; larger window sizes will highlight semantic concepts and other relationships that hold over larger scales. » [*ibid.* : 23]

D'après les auteurs, cela peut s'expliquer par la variance de la distance entre deux mots : les mots d'une expression figée comme *bread and butter* auront une variance quasi-nulle, alors que la distance entre des « concepts » comme *homme* et *femme* est beaucoup plus importante. Une large

16 http://www.oucs.ox.ac.uk/rts/xaira/
17 http://www.lexically.net/wordsmith/
18 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

fenêtre servira donc mieux nos intérêts. Les auteurs utilisent l'Information Mutuelle pour ordonner les collocations, ce qui leur permet de donner un plus fort score aux associations significatives. Pour deux mots x et y, leur information mutuelle I(x, y) correspond au rapport entre la probabilité de les identifier ensemble et la probabilité de les trouver indépendamment :

$$I(x,y)=\log_2 \frac{P(x,y)}{P(x).P(y)}$$

Les probabilités sont estimées en normalisant la fréquence sur la taille du corpus. L'estimation de probabilités nécessite de larges corpus (dizaines voire centaines de millions de mots) ce qui n'est pas notre cas : nous présenterons toutes les collocations que nous étudions en contexte.

Le tableau (3.2) fait figurer les collocats nominaux du verbe *abandonner* (fréquence absolue f=11) extraits à partir de ces conditions :

Forme	Fréquence
cabane	2
château	1
dernier	1
dispute	1
jour	1
maman	1
pays	1
pénombre	1
privilège	1
projet	1
terre	1

Tableau 3.2 – Collocats nominaux du verbe abandonner

Comment interpréter une telle liste ? On peut dans un premier temps réunir les mots en fonction d'affinités sémantiques afin de généraliser les relations sémantiques entre verbe et nom. Par exemple, le [Lieu], qui concerne *cabane*, *château*, *pays* et *terre*, semble être la catégorie dominante. Le verbe *abandonner* serait sémantiquement lié à la notion de localisation. On pourra évidemment proposer d'autres catégories.

On peut dans un second temps, chercher à valider la pertinence de cette catégorie : *cabane*, *château*, *pays* et *terre* sont-ils toujours employés comme lieux, les verbes avec lesquels ils se combinent sont-ils toujours des verbes de localisation ? En d'autres termes, ces catégories sont-elles valides à plus large échelle ? En analysant les contextes partagés par ces quatre noms, on trouve que 8 verbes sont des collocats de ces quatre noms (tableau 3.3), 14 uniquement pour trois (tableau 3.4), et 22 pour deux (tableau 3.5).

verbe	cabane	pays	château	terre	Total
être	7	9	8	19	43
avoir	6	3	2	10	21
arriver	2	2	10	2	16
faire	3	2	1	2	8
voir	1	1	3	2	7
trouver	2	1	1	1	5
pouvoir	1	1	1	2	5
abandonner	1	1	1	1	4

Tableau 3.3 – Verbes communs aux 4 noms

verbe	cabane	château	terre	pays	Total
aller		2	4	4	10
habiter	2	5		1	8
venir		1	2	4	7
entrer	2	3	1		6
vivre	2	2		2	6
partir		1	1	4	6
revenir		2	1	1	4
approcher	2	1	1		4
dire	2	1		1	4
quitter		1	1	2	4
retrouver	1	1	2		4
vouloir	1		2	1	4
donner		1	1	1	3
rentrer	1	1		1	3

Tableau 3.4 – Verbes communs à trois des noms

verbe	cabane	château	terre	pays	Total
poser		2	6		8
retourner	1	6			7
construire		5	2		7
sortir	1	4			5
rendre		3	1		4
mettre		2	2		4
prendre	1	2			3
rester		2	1		3
cacher		1	2		3
filer		1	2		3
installer	1		2		3
apercevoir		1	1		2
crier	1	1			2
décider		1		1	2
disparaître	1	1			2
entendre		1	1		2
marcher		1	1		2
regagner	1	1			2
rencontrer		1	1		2
aimer	1		1		2
parler	1		1		2

Tableau 3.5 – Verbes communs à deux des noms

À partir de ces tableaux de cooccurrence, il est possible de tirer plusieurs conclusions. Tout d'abord, les cooccurrences qui sont partagées par tous les noms ne sont pas nécessairement les plus sémantiquement informatives : les verbes *être*, *avoir*, *pouvoir* et *faire*, par exemple, sont généralement fréquents en corpus et, de fait, constituent un indice peu informatif pour identifier des propriétés partagées. Par opposition, le verbe *arriver* semble confirmer la notion de localisation.
Dans les deux autres tableaux, on observe qu'un grand nombre de verbes partagent la notion de localisation ou de déplacement, ce qui conforte la première analyse et en appelle une seconde : on peut alors proposer des catégories sémantiques à partir des affinités entre verbes (et non plus entre noms), en ne se limitant plus uniquement au lieu, mais en affinant cette catégorie et en en proposant d'autres. Les rapprochements que l'on peut opérer sont de l'ordre de la fonction sémantique remplie par les noms vis-à-vis de la classe de verbe. Par exemple, on peut mettre en lumière les propriétés résumées dans le tableau 3.6.

CATÉGORIE	Exemples de verbe
[DESTINATION]	*trouver, arriver, retrouver, approcher, entrer, rentrer, revenir, aller, venir, (se) rendre, regagner, filer, retourner, marcher*
[POINT DE DÉPART]	*abandonner, quitter, partir, sortir*
[DOMICILE]	*habiter, vivre, rester, cacher, installer*
[OBJET DE VISION]	*voir, disparaître, apercevoir*
[LIEU D'ACTIVITÉ]	*chasser, rencontrer, entendre*
[MOYEN D'ÉCHANGE]	*donner, prendre*
[ÉDIFICE]	*construire*

Tableau 3.6 – Catégories sémantiques associées aux collocats verbaux

Les catégories que nous avons identifiées ne sont pas toutes liées à la localisation : les verbes *voir* et *disparaître* partagent la notion de *visibilité* et les quatre noms peuvent être associés aux verbes *donner* et *prendre* dans des contextes d'échanges. Le reste des catégories est associé à la localisation à différents niveaux de granularité : la destination, le point de départ, le domicile, le lieu d'activité et l'édifice. Remarquons que le verbe *abandonner* fait désormais partie d'une plus grande classe qui catégorise les noms comme *cabane* en [Point de Départ].

Cette catégorisation des noms en fonction des collocats verbaux comporte un danger : elle est effectuée intuitivement sans avoir observé les concordances (le co-texte) et peut à ce titre être contredite par les données. Par exemple, le classement du verbe *venir* dans [Destination] peut être justifié dans des structures du type */venir à Nom/*, mais non */venir de Nom/*. Prendre en compte la nature des prépositions dans de tels cas implique de dépasser la limite d'une analyse collocationnelle binaire. Une analyse sémantique complète exige d'observer les concordances pour valider empiriquement ces regroupements. Ceci nous permettra également de décortiquer le fonctionnement de l'analyse collocationnelle.

3.2.2. Analyse des concordances

Prenons l'exemple de la catégorie [Destination]. Voici les concordances liées au verbe *trouver* (la taille de la fenêtre est signalée en italique) :

(25)*Il courut vite voir, pour* trouver *Philia assise par terre, avec* un lynx mort à coté d' elle
(26)Cachette est une petite souris *brune-grise que Zazounette a* trouvée *dans sa cabane en haut* d'un arbre
(27)-*Mais où se* trouve *le pays froid* ?
(28)Issa *entra dans le château espérant* trouver *de l' aide*
(29)*Arrivé à la cabane il* trouva *Fanchon en train de soigner* leurs trois maigres poules
(30)Le créateur qui avait entendu les trois spiritueux faire leurs bruits terribles, entendit aussi les personnes *sur terre pleurant dehors pour* trouver *de l' aide*.

L'exemple (27) est une occurrence de construction pronominale (prédicat différent) qui modifie le sens du verbe : ce n'est plus la destination (*trouver un lieu*) mais le domicile qui est pertinent. L'exemple (30) est un cas d'association fortuite : il n'y a aucun rapport entre *terre* et *trouver* puisque *terre* sert à localiser *personnes* ; notons que dans ce dernier cas, *terre* est synonyme de planète, alors qu'il signifie *sol* en (25). En règle générale, si l'on exclut les exemples (27) et (30), le nom de la catégorie que nous avons nommé [Lieu], agit comme un cadre dans lequel un protagoniste trouve quelque chose. Il s'agit donc de la *localisation* du procès (« trouver quelque chose dans un lieu ») et non pas, comme nous l'avions classé, d'une destination (dans le sens « trouver un lieu », comme

objet direct du verbe).

Le verbe *arriver* apparaît dans les concordances suivantes :

(31)*Le gardien du château* arrive *en disant : -Prince, vous avez* une visite, c' est la princesse Flora"
(32)Ce matin -là, un peu avant la cloche de *neuf heures cette fois, Linou* arriva *au château*
(33)*Elle* arrive *au château*
(34)*Le maître chat* arriva *enfin dans un beau château* dont le maître était un ogre, le plus riche qu'on ait jamais vu, car toutes les terres par où le roi avait passé étaient sous la dépendance de ce château
(35)Ils quittent sans regret *les terres du Marquis et* arrivent *à la maison*
(36)Arrivé *au pays de Zazounette, Pier-Paulet*, Catherinette, Martinette, Hélènette et tous les autres accueillent Zazounette
(37)Notre ami, vêtu de l' armure, *franchit le volcan et bientôt* arrive *près d'un château*
(38)Arrivé *à la cabane il trouva* Fanchon en train de soigner leurs trois maigres poules
(39)Le clown put *poursuivre son chemin mais en* arrivant *au château il vit la* sorcière qui l'avait ensorcelé et préféra ne pas entrer
(40)*Après une demi-heure il* arrive *au château, mais ne voit* pas le boulet !
(41)*Enfin, ils* arrivent *près de la cabane du* grand sorcier
(42)Aujourd'hui Zazounette, *Cachette, Chucky, Canellette et Samour* arrivent *au pays lointain (des "ie"* et des "io")
(43)*En* arrivant *au château , les serviteurs entendent* le bruit des chevaux
(44)Le lendemain , après avoir *bien dormi , une Linou rayonnante* arriva *au château*
(45)*Olivier* arriva *au château* .

Pour ce verbe, l'intuition a plus de réussite puisque seul l'exemple (31) ne répond pas à la catégorie [Destination], dans lequel le nom n'entretient aucun rapport avec le verbe. En contrôlant l'ordre des mots (à droite), on peut écarter ce cas. Deux avantages d'une méthode basée sur les collocations apparaissent également dans les concordances : d'une part, la taille de la fenêtre (5 mots) permet de capter des relations pertinentes pour des mots qui sont à des distances variables du mot-cible (de 2 pour l'exemple 43, à 5 pour l'exemple 34) ; d'autre part, la variation de la préposition intermédiaire (*près de, à, dans*) n'a aucune influence sur la catégorie [Destination].
Voyons à présent des verbes associés aux membres de la catégorie [Point de départ] comme *abandonner* et *quitter*. Voici les concordances pour chacun d'eux successivement :

(46)Dépitée, la baronne *Sancœur quitta définitivement le pays*, abandonnant *terres et privilèges*
(47)*Elle vivait dans une cabane* abandonnée, *délabrée, faite de murs* en torchis
(48)Il est vert ; *il habite dans un château* abandonné
(49)Ils virent au *loin une cabane qui semblait* abandonnée
(50)-Tout se passe comme il faut ; voici maintenant pour toi *un extraordinaire déguisement pour* quitter *le château dès cette nuit*
(51)*Dépitée, la baronne Sancœur* quitta *définitivement le pays, abandonnant terres* et privilèges
(52)*Ils* quittent *sans regret les terres du* Marquis et arrivent à la maison
(53)Linford le lit tout *haut : "Chers amis, nous vous* quittons *pour rejoindre le pays du* lac à la douce mélodie

Si le verbe *abandonner* peut certes impliquer que quelque chose ou quelqu'un qui y résidait a décidé de le quitter, le verbe semble plutôt mettre l'accent sur la dépopulation du lieu, ce qui peut être imputé à sa forme participiale. Il s'agit donc d'une qualité d'un lieu que d'être peuplé ou dépeuplé

(dépeuplement causé par un départ) que la prise en compte d'informations syntaxiques permettrait de distinguer. Concernant le verbe *quitter*, c'est bien le point de départ qui domine sauf en (53).

Observons à présent les contextes des verbes *vivre* et *cacher* dans la catégorie [Domicile].

(54)Près de *ce petit reste de château*, vivait *un village de lézards*
(55)*Peau d'âne* vivait *donc seule dans cette cabane* un peu à l'écart de la ferme que lui avait attribuée la patronne du lieu pour qu'elle n'empeste pas et n'incommode pas les autres avec sa crasse et son odeur de peau d' âne
(56)*Cependant le roi, qui* vit *en passant le beau château* de l'ogre, voulut y entrer
(57)*Elle* vivait *dans une cabane abandonnée, délabrée*, faite de murs en torchis
(58)*Ils* vécurent *ainsi dans ce pays beau* et chaud, très longtemps et heureux ensemble
(59)Notre vengeance frappera les criminels *et les pays où ils* vivent
(60)Le livre du vieil explorateur donnait bien *le nom du pays où* vivait *le Cracodile, Chiméria, mais ne* donnait pas grands détails
(61)Le *ver de terre, effrayé, se* cachait *derrière la rose*
(62)Lui qui était si fier de sa queue en panache quand il vit sa queue aussi nue qu'un ver *de terre , humilié il la* cacha *entre ses pattes arrière, et* s'en alla ainsi sans demander son reste
(63)-Puisque mon vœu s'est réalisé, je vais *vous l'avouer : il est* caché *dans la bibliothèque du château* des Cambouts de Coislin.

L'exemple (56) est un cas intéressant (et courant) d'erreur d'analyse morphosyntaxique due à une homonymie de la forme *vit* (verbe *voir* ou *vivre*) : sans une analyse du contexte, il est impossible de se prononcer sur son sens. Si le lemme avait été correctement identifié, la relation [Objet de vision] aurait été valable.
Concernant le verbe *cacher*, seul l'exemple (63) reste approprié à cette catégorie (il peut néanmoins être discuté) : dans les deux autres exemples, le nom *terre* fait partie d'un mot composé, *ver de terre*, ce qui fausse l'analyse. On remarque enfin que ce verbe s'emploie à la forme pronominale (61) à la voix active (62) et passive (63) et que pour tous ces cas, il y a bien un nom dont la fonction est [Domicile] (*rose*, *pattes*).

Pour les verbes *voir* et *disparaître* de la catégorie [Objet de vision], on obtient les concordances suivantes :

(64)C'est alors *qu'il se rappelle avoir* vu *dans les ruines du château* d'Aubigné, un village voisin, un énorme boulet de canon !
(65)Ils partent ensemble dans la grande *montagne d'où l'on* voit *le pays de Zazounette*
(66)Pauline et Xavier, Cécile et David, se précipitèrent *hors de la cabane et* virent *en effet un gros cheval* marron, beaucoup moins fin qu'un pur-sang, tout sellé, qui paissait tranquillement l'herbe de la prairie de Mamie
(67)Au nord de *ce village l'on pouvait* voir *un immense et sombre château*
(68)Le maître chat arriva enfin dans un beau château dont le maître était un ogre, le plus *riche qu'on ait jamais* vu, *car toutes les terres par* où le roi avait passé étaient sous la dépendance de ce château
(69)Le clown put poursuivre son chemin mais *en arrivant au château il* vit *la sorcière qui l'avait* ensorcelé et préféra ne pas entrer
(70)Et la princesse, confiante, retourna *vers le château et alla* voir *son père pour lui demander* cette fois la robe couleur soleil
(71)Après une demi-heure il *arrive au château, mais ne* voit *pas le boulet !*
(72)Le *ver de terre avait tout* vu *; tremblant de peur, il expliqua* que la taupe n'avait pas fait exprès : "J'ai été poursuivi par une taupe ; heureusement qu'elle a perdu ses lunettes !

(73)Un jour, il se *rendit au château pour le* voir
(74)*Ils* virent *au loin une cabane qui* semblait abandonnée
(75)*Il regarde par terre et* voit *des bêtes, des mille-pattes, des* serpents et des araignées
(76)Etant donné que les brigands voyaient à l'avance qui s'approchait *de leur cabane, le prince* disparut *à l'aide de son* épée
(77)*Aussitôt, tous les habitants* disparurent*, le château s'enfonça et* il ne resta plus qu'une bosse.

Pour le verbe *voir*, les noms alternent entre la catégorie [Lieu d'activité] (64, 66, 69, 70, 71, 73, 75) et [Objet de vision] (65, 67, 74), soit une double catégorisation vis-à-vis de ce verbe. L'utilisation de la relation syntaxique Objet direct permettrait d'isoler la catégorie [Objet de vision] ; en revanche, la relation avec [Lieu d'activité] (interprété comme cadre de localisation) ne correspond pas toujours à une relation syntaxique. On retrouve l'erreur du *ver de terre* (72), ainsi que celle d'association fortuite en (68). Le verbe *disparaître* (76, 77) n'entretient aucune relation avec ses collocats.

Les collocats des verbes *chasser* et *rencontrer* ont été classés comme [Lieu d'activité] dans le sens de « lieu où se déroulent des événements comme les rencontres ou la chasse » (comme nous avons vu, nous l'avons élargi à l'idée de cadre de localisation). En voici les concordances :

(78)Un jour, en allant *chez une copine, la taupe* rencontra *un ver de terre et* voulut le manger
(79)Il y a *longtemps, un prince du château* rencontra *une belle princesse de la* famille de Flora.
(80)Bien sûr la princesse ne voulut pas d'un casse noisette *comme mari, alors on le* chassa *du château*
(81)Vous avez fait une grosse sottise de trop, et je vous chasse de la *terre*
(82)Le Roi des lutins s'est fâché très fort et leur a dit que s'ils faisaient encore une bêtise *de cette taille, ils seraient* chassés *de la terre*.

Le verbe *rencontrer* n'entretient aucune relation avec ses collocats : l'association est fortuite. Le problème pour *chasser* est qu'il ne s'agit pas du sens que nous avions prévu : il est ici employé pour signifier « faire quitter quelqu'un d'un endroit » et non de « poursuivre pour capturer ». Il relève alors plutôt de la catégorie [Point de départ]. Remarquons que ce verbe met en jeu trois entités (sens causatif) : l'entité « chassée » est celle qui quitte le lieu et ce départ est provoqué par une troisième entité (*on* en 80 et *je* en 81).

Les collocats des verbes *prendre* et *donner* ont été classés par la catégorie [Moyen d'échange] :

(83)Tu *peux entrer dans le château,* prends *le couloir et la première* porte à droite
(84)Les jeunes *gens conduits par la fée* prennent *le chemin du château*
(85)Sur le chemin du *retour au château, Linou put* prendre *son temps : tout le travail* était fait
(86)*Très vite, la cabane* prit *un air de fête avec* tout bien installé sur la table en rondin, et les enfants s'assirent autour sur les bûches qui servent de sièges
(87)Au bout d'un *certain temps, la cabane avait* pris *un air coquet, comme si* elle était habitée depuis plusieurs jours, et ils étaient très fatigués tous les quatre
(88)Deux chemins sont inondés : l'un permettant d'aller à Saint Rémi *du Plain et l'autre* donnant *l'accès aux terres du* Marquis
(89)*Je vous* donnerai *là-bas des châteaux et des* terres
(90)*Le livre du vieil explorateur* donnait *bien le nom du pays* où vivait le Cracodile, Chiméria, mais ne donnait pas grands détails

Les verbes *prendre* et *donner* sont très fréquents dans le corpus, et, comme le prédit le principe d'idiome de Sinclair (cf. *infra* 2.3.3), de nombreuses occurrences font partie de locutions verbales : *prendre le chemin*, ou *donner (l')accès à*. Cette contrainte idiomatique s'accompagne d'un changement de sens qui entraine un mélange de contextes ayant peu de rapport. En fin de compte, aucun exemple ne semble véritablement exploitable et la catégorie [Moyen d'échange] ne s'applique pas à ces collocats. Pourtant, le sens d'échange est bien le premier qui nous vient à l'esprit pour définir ces verbes (on pourra s'en convaincre en consultant un dictionnaire), comme en (89).

Pour finir, voyons le verbe *construire* pour la catégorie [Édifice] :

(91)Revenant de croisade, le seigneur, épuisé mais heureux *voulut récupérer ses terres et* construire *son château à l'endroit* où habitait la sorcière
(92)-Le Seigneur *veut détruire ces bois pour* construire *un château plus beau que* celui du Matz
(93)Le seigneur *fit venir des ouvriers qui* construisirent *le château avec quatre grandes* tours reliées par des remparts
(94)Sept mots y *étaient écrits : "Au secours, Seigneur* construire *château dans bois"*
(95)Pendant ce temps-là les enfants s'*organisaient dans le souterrain : ils* construisaient *des objets en terre, se* nourrissaient avec le lait des vaches et les œufs des poules qu'ils avaient réussi à sauver
(96)Acorto arrive dans la *cour de l'étrange château* construit *sur une colline déserte*

On observe que la notion d'édifice ne s'applique qu'aux châteaux. Pour le nom *terre* (les autres noms ne sont pas en collocation avec ces verbes), il s'agit du matériau (95) ou du [Lieu d'activité] (91).

3.2.3. Synthèse

Nous avons récapitulé ces relations sémantiques dans un réseau entre verbes (en orange) et noms (violet), reliés par des relations nommées selon les catégories identifiées et corrigées (Figure 3.2).

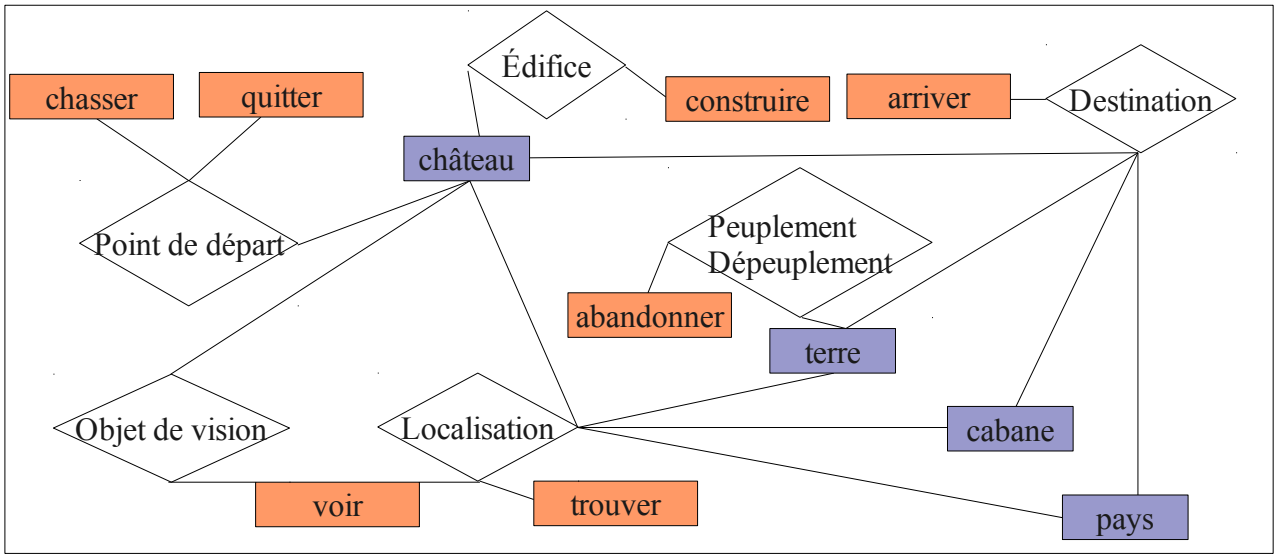


FIG 3.2 – Réseau sémantique de collocats verbe-nom

Sur ce schéma ne figurent évidemment pas toutes les collocations analysées ni toutes les catégories possibles, ni tous les verbes ; il est donné seulement à titre d'illustration. Cette visualisation permet de mieux saisir ce que peut désigner une catégorie dans le contexte d'une analyse collocationnelle : elle regroupe un ou plusieurs mots-types (ici des lemmes) qui ont un comportement collocationnel similaire. Cette méthode permet d'opérer une double classification, celle des noms et celle des verbes, par l'intermédiaire des catégories. La catégorie étiquette le rôle que joue le nom dans la scène que le verbe décrit, ou sa fonction, qui de la même manière permet de définir une facette du nom. Les classes de verbes sont créées en fonction des rôles qu'ils partagent. Par exemple *chasser* et *quitter* font partie d'une même classe étant donné ce critère, possédant tous deux un rôle qui désigne un « point de départ ». Néanmoins, comme nous l'avons noté, ces deux verbes diffèrent du fait que l'un est causatif et met en jeu trois entités (un « chasseur », un « chassé » et un « point de départ ») alors que l'autre ne met en scène qu'un « quitteur » et un « point de départ ». À titre de remarque, la méthode permet évidemment de classer le verbe qui constituait au départ notre point d'entrée (*abandonner* est associé à la catégorie [Peuplement/Dépeuplement]). Enfin, on peut observer que les noms entrent régulièrement dans plusieurs relations (*château* en recense 5).

À travers cette courte étude, nous avons pu mettre à jour les intérêts mais également certaines limites de l'analyse collocationnelle traditionnelle (et de son automatisation). Tout d'abord, la coocurrence d'un verbe et d'un nom peut être fortuite car ils n'entretiennent aucune relation contextuelle. Deux cas peuvent se présenter :

- Lorsque le collocat fait partie d'une unité plus large dont il n'est pas la tête (mot composé comme *ver de terre* ; syntagme comme dans *le gardien du château arrive*)
- Lorsque le collocat n'est lié au mot-cible que par la proximité (exemples 76 et 83) parce qu'ils font partie de deux propositions différentes.

Ensuite, la polysémie rencontrée demande un meilleur contrôle du sens en s'appuyant sur des indices contextuels supplémentaires. Pour dépasser ce problème, il est possible :

- d'explorer l'impact de la position du collocat (position gauche/droite et distance) dans la fenêtre sur les regroupements obtenus pour ne pas mélanger les collocats apparaissant dans différentes positions d'une fenêtre,
- de dépasser une vision atomiste (ou binaire) du sens en constituant des patrons prenant en charge un plus grand nombre d'éléments contextuels,
- et de se doter d'informations plus riches sur les données : catégories morphosyntaxiques, syntaxiques, ontologiques, etc.

L'analyse collocationnelle offre un cadre d'analyse qui peut constituer un atout si l'on contrôle les différents paramètres en jeu. Cela suppose la conception d'un environnement de travail plus adapté et personnalisé qu'un simple concordancier [Manning, 2003 : 292–293].

3.3. Des patrons collocationnels syntaxiques

Pour dépasser une vision atomiste des relations sémantiques, on peut s'appuyer sur l'analyse syntaxique. Elle permet de structurer l'environnement d'un mot-cible : les mots sont regroupés en constituants et les relations, ou dépendances de ces constituants sont identifiées. La résolution des constituants (ou encore « chunking ») consiste à déterminer les frontières des syntagmes en fonction de leur catégorie morpho-syntaxiques (syntagme nominal SN, verbal SV, adjectival SA, adverbial SAdv, etc.) et de leur réalisations syntaxiques (pronom, syntagme nominal, infinitive, etc.). L'identification des relations de dépendance correspond aux relations entretenues entre le syntagme verbal et les syntagmes qui l'environnent (sujets, objets, circonstant, etc.) ainsi que les relations entretenues par les éléments au sein d'un syntagme (spécifieur, modifieur, auxiliaire, etc.). Les relations syntaxiques qui nous intéresseront ici sont les relations entre syntagmes.

Il existe un certain nombre d'analyseurs syntaxiques automatiques en dépendance pour le français, comme Syntex [Bourigault, 2007]. Ces derniers facilitent grandement le travail d'annotation syntaxique, mais n'atteignent pas la qualité d'une annotation manuelle. Il faut également savoir que de tels systèmes implémentent des modèles théoriques et effectuent ainsi des choix qui peuvent être discutables. Il en va de même pour les ressources comme les lexiques syntaxiques que nous discuterons en 3.3.2. Dans notre cas, nous avons tenu à réaliser une annotation manuelle pour les raisons suivantes :

- Nous rendre compte par nous-mêmes des difficultés dans l'analyse syntaxique automatique et dans l'application de conventions d'annotation syntaxique.
- Disposer d'une base fiable pour l'analyse sémantique
- Guider l'annotation par des principes sémantiques

Nous avons ainsi enrichi le corpus de contes d'annotations syntaxiques.

3.3.1. Extraction manuelle de relations syntaxiques

La première précaution à suivre pour effectuer une annotation manuelle du corpus est la numérotation des mots du corpus pour conserver un lien entre la ressource et le texte (autorisant ainsi un retour ultérieur) et permettre de travailler sur un fichier d'annotation séparé. Nous nous sommes concentré sur les 90 verbes les plus fréquents du corpus, qui figurent dans le tableau 3.7.

Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence
dire	731	commencer	121	habiter	73	écouter	49	se promener	34
savoir	290	crier	116	connaître	73	pousser	49	montrer	33
venir	286	s'appeler	114	comprendre	72	se coucher	49	se asseoir	33
prendre	265	attendre	113	se approcher	72	boire	48	chanter	33
trouver	248	revenir	113	retourner	70	se réveiller	48	bouger	32
partir	224	appeler	111	tirer	70	arrêter	48	couper	31
demander	216	parler	102	finir	69	tenir	46	sourire	30
falloir	202	croire	101	marcher	69	réussir	44	se reposer	28
donner	191	aimer	101	jouer	64	avancer	42	éteindre	18
regarder	187	laisser	95	se lever	62	pleurer	42	battre	17
passer	180	courir	95	dormir	61	rire	41		
entendre	177	aider	95	préparer	59	conduire	41		
sortir	168	rentrer	90	cacher	58	jeter	40		
manger	158	ouvrir	87	perdre	54	emmener	40		
rester	149	monter	87	porter	54	lancer	39		
chercher	140	raconter	86	apporter	54	attraper	39		
décider	135	sauter	80	descendre	51	sentir	38		
devenir	129	retrouver	79	rendre	50	tuer	36		
penser	123	essayer	77	poser	50	servir	35		
tomber	121	voler	73	oublier	50	tourner	35		

Tableau 3.7 – Liste de fréquence des verbes du corpus de contes

Comme on peut le constater, les formes pronominales sont distinguées et les auxiliaires et modaux ont été exclus ; ces derniers agissent plutôt comme auxiliaires du syntagme verbal que véritablement comme mise en relation de syntagmes nominaux.

Le format adopté pour l'annotation syntaxique est un triplet *<REL, V, A>* où *REL* correspond à la nature de la relation et *V* et *A* sont respectivement les identificateurs (numéros de mots) du verbe et de son argument (voir tableau 3.8). Nous n'avons qu'annoté la tête du verbe et de l'argument. Le jeu de relations obtenu à la fin de l'annotation est résumé dans le tableau 3.8 :

RELATION	Fréq.	Exemples
SUJ	11006	Oui, je sais
OBJ	5328	Fouad s'est fait arrêté
IOBJ	805	il les menaça pourtant de leur faire couper la tête
INF	666	Les enfants, courez faire sonner le tocsin une nouvelle fois
IMPERS	378	il suffit d'y croire
QUAL	327	Les pieds de Jérôme sont devenus trop grands pour nous
LOC	288	Il faut que vous habitiez ailleurs
TPS	273	Le lendemain, Alice se leva très tard
MAN	193	Il n'était plus capable de marcher droit
NAME	158	Salut, je m'appelle Max
NEG	110	Personne ne le savait
OBJQUAL	103	Franchement, Belinda, je trouve tout cela inquiétant
IDIOM	65	On voit mal comment il s' y prendrait
ABOUT	56	elle en tenait une boîte à la main
QUANT	42	elle en avait trouvé 234
OBJQUANT	28	et voilà Loup-Bleu qui s'élance pour en boire une gorgée
PROV	16	je vais retourner là-bas

Tableau 3.8 – Liste des relations syntaxiques les plus fréquentes

La relation Sujet est la plus fréquente et l'exemple correspond au triplet *<REL=SUJ, V=savoir, A=je>* (ce sont des numéros de mots plus exactement). Sa détection couvre en réalité des phénomènes syntaxico-sémantiques complexes. Tout d'abord, elle est distinguée des sujets impersonnels (relation IMPERS). La relation SUJ suppose donc l'identification du sujet logique (et non uniquement grammatical). Par exemple, un verbe à l'infinitif n'a pas nécessairement de sujet syntaxique explicite : le sujet logique peut se trouver dans la proposition matrice, voire plus loin. Notre étude étant guidée par des principes sémantiques, nous avons préféré, dans la mesure du possible, identifier exhaustivement le mot désignant le référent en relation avec le verbe, même s'il n'était pas lié par une relation syntaxique directe au verbe. Par exemple, nous avons considéré que le sujet logique du verbe *manger* dans l'exemple (97) était *nous* :

(97)Dépêche toi, lui disait Philia, il nous reste juste le temps de manger quelque chose avant de partir.

Ajoutons la difficulté liée au mode impératif (par exemple *donne ça !* ; voir aussi le verbe *prendre* dans l'exemple 99) pour la détection du sujet : le sujet « fait partie » du verbe. Les impératifs ont été pris en compte en indiquant le même identifiant pour le verbe et l'argument, ce que nous exploiterons par la suite (cf. *supra* 4.1.1).

L'exemple de la relation OBJ illustre également les divergences entre relations syntaxiques et sémantiques : le SV *s'est fait arrêté* est complexe (pronominal et causatif) et la relation entretenue entre les *têtes* du SN et du SV est une relation Objet (*X arrête Y*) alors que la relation entre ces deux syntagmes est de type Sujet. De la même manière, nous avons résolu manuellement la forme passive et considéré les compléments d'agent comme sujet.

Dans certains cas il n'est pas possible d'identifier sans ambiguïté la tête du syntagme. Par exemple, nous avons souhaité mettre en relation un verbe de parole et un segment de discours direct (entre guillemets ou pas), comme illustré en (98).

(98)" Pousse-toi, triple idiot ! " dit Jennyfer.

Cette relation n'est en général pas considérée comme syntaxique, et les analyseurs automatiques ne les détectent pas. Pourtant d'un point de vue sémantique, on peut la considérer comme une variante de la structure */X dire que Proposition/*, que nous avons prise en compte. C'est pour éviter des incohérences que nous avons inclus la totalité de l'empan comme complément du verbe.

Les coordinations d'arguments sont un autre exemple de difficultés posées par l'identification d'une tête unique comme argument d'un verbe. Pour résoudre cette difficulté, nous avons dupliqué la relation de dépendance : il y a donc deux relations objet au verbe prendre dans l'exemple (99).

(99)Tu peux entrer dans le château , prends le couloir et la première porte à droite

Lorsqu'un complément du verbe est un syntagme prépositionnel (SP), ou une proposition complétive (introduite par *que* ou *qui*), la forme de la préposition ou de la conjonction devient l'étiquette de la relation (par exemple *<REL=dans V=... A=...>*).

La relation Objet indirect (IOBJ) a été choisie pour regrouper les pronoms indirects uniquement (*leur* dans le tableau 3.8).

Enfin, l'analyse syntaxique traditionnelle distingue les compléments obligatoires des compléments optionnels (arguments/circonstants), mais en pratique, il est difficile de se prononcer sur cette distinction. Par exemple, un adverbe de temps peut être jugé tantôt facultatif, tantôt essentiel.

(100)Mais nous ferions mieux de manger vite et d'aller montrer les signatures.

Dans l'exemple (100), le verbe ne met certes pas en relation une entité comme dans *manger une pomme*, mais le sens du verbe est tout de même affecté : l'adverbe *vite* permet de focaliser sur un aspect de l'activité *manger*., dans ce cas, la manière (MAN). Les relations LOC (pour lieu), TPS (pour temps) et PROV (pour provenance) sont du même ordre. Elles se justifient par la présence d'adverbes (*ici*, *là-bas*) ou de pronoms (*en*, *y*), dont la référence est imprécise : la relation apporte une information supplémentaire sur la nature de la contribution de ces mots.

3.3.2. Exemple de patron syntaxique

Une fois les relations syntaxiques d'un verbe repérées, il est possible de restituer pour chaque occurrence le patron syntaxique, qui correspond à la somme des relations syntaxiques entretenues par une occurrence verbale. Pour exemple, le verbe *répondre* (f=165) se présente dans 22 constructions différentes (tableau 3.9) et jamais en forme pronominale.

Constructions syntaxiques	Fréq.	Prop.
[[SUJ]] [[SpeechAct]]	83	50%
[[SUJ]] [[IOBJ]] [[SpeechAct]]	23	14%
[[SUJ]] [[IOBJ]]	9	5%
[[SUJ]]	9	5%
[[SUJ]] [[à]]	7	4%
[[SUJ]] [[OBJ]]	6	4%
[[SUJ]] [[IOBJ]] [[que]]	6	4%
[[SUJ]] [[MAN]] [[SpeechAct]]	4	2%
[[SUJ]] [[en]] [[SpeechAct]]	3	2%
[[SUJ]] [[avec]] [[SpeechAct]]	2	1%
[[SUJ]] [[à]] [[SpeechAct]]	2	1%
[[SUJ]] [[SpeechAct]	1	1%
[[à]] [[par]]	1	1%
[[SUJ]] [[IOBJ]] [[OBJ]]	1	1%
[[SUJ]] [[à]] [[TPS]]	1	1%
[[SUJ]] [[en]] [[IOBJ]] [[SpeechAct]]	1	1%
[[SUJ]] [[IOBJ]] [[MAN]] [[SpeechAct]]	1	1%
[[SUJ]] [[en]] [[IOBJ]]	1	1%
[[SUJ]] [[que]]	1	1%
[[SUJ]] [[IOBJ]] [[MAN]]	1	1%
[[SUJ]] [[à]] [[en]] [[TPS]]	1	1%
[[SUJ]] [[IOBJ]] [[par]]	1	1%
Total	165	100%

Tableau 3.9 – Patrons syntaxiques du verbe répondre
Fréq.=fréquence ; Prop.=proportion

On constate tout d'abord que la grande variété des patrons syntaxiques s'explique par la présence/absence d'arguments, et ce, malgré le fait que les contraintes d'ordre des arguments aient été dissoutes (exemples 101 et 102) et que les complétives en que aient été associées aux segments de discours direct (relation [[SpeechAct]]).

(101)À cela le rapiat répondit -Pas de sentiments, ma mignonne, c'est comme ça qu'on perd de l' argent !
(102)-[…] écarte-toi de mon chemin, si tu ne veux pas perdre ton panache, répondit le rat au tigre.

Ce type d'alternance (l'absence/présence des arguments) n'a aucune influence sur le sens du verbe. Il conviendrait donc de les associer : le verbe *répondre* est principalement employé pour rapporter des discours : 72% des occurrences contiennent la catégorie [[SpeechAct]] (103), qui désigne un discours rapporté.

(103)Pénélope lui répond en riant : "Oui ... et nous aussi ! ! "
Patron : [[SUJ]] [[IOBJ]] [[en]] [[SpeechAct]]

Si le sens du verbe ne semble pas changer, les arguments relevant de la même relation syntaxique peuvent jouer différents rôles et être de nature différente.

Lorsqu'un complément de ce type (relations [[OBJ]], [[que]] ou [[SpeechAct]]) est absent, le verbe se focalise sur le destinataire ([[IOBJ]], [[à]] ; 104), ou sur ses paroles ([[à]] ; 105).

(104)Tout en répondant au chef, le diamant se désintègre dans la main de Fouad, sous le regard inquiet du chef de la bande
(105)Réponds à ma charade et tu pourras entrer.

On observe que ces deux catégories sémantiques sont réalisées par les mêmes moyens syntaxiques (syntagme prépositionnel introduit par la préposition *à*), mais que ces contraintes syntaxiques ne suffisent pas pour distinguer le destinataire (*chef*) de ses paroles (*charade*). La situation se complique lorsque l'on compare le nom *charade* à *désir* en (106) ou *violence* en (107).

(106)Mais sa mère Rose ne répondait pas à ses désirs .
(107)Ces hommes qui ont péri au nom de la vengeance vous supplient de ne pas répondre à la violence faite à votre peuple par la vengeance

Même en présence d'un destinataire identifié, un sujet inhabituel peut être employé pour traduire une atmosphère sonore, suggérant implicitement l'absence de réponse (108-109).

(108)-Est-ce toi le ciel qui me parle ? Et seul le vent lui répondit.
(109)Seul un éternuement épouvantable répondit au Père Noël.

Enfin, on trouve des formes intransitives, comme en (110).

(110)L'une d'elles a téléphoné à Alexandre. Il n'était pas là et Louise, sa femme, a répondu, mais elle n'a pas de voiture.

Il s'agit ici d'un sens spécialisé de *répondre*, que l'on peut interpréter comme la forme elliptique de *répondre au téléphone*, dans laquelle le SP en *à* est un moyen de communication. Mais la forme intransitive n'est pas garante d'un tel usage (111).

(111)Fouad ne répond pas parce qu'il n' a pas la réponse

On peut croire que la prise en compte de la négation permettrait de les distinguer.

3.3.3. Comparaison de ressources

Si l'on compare les patrons obtenus avec une des ressources de cadres syntaxiques comme Lexschem (dont la méthodologie d'extraction est décrite dans [Messiant, 2009]), établi à partir des analyses de Syntex sur un corpus de presse, on constate que l'argument syntaxique majeur que nous avons mis à jour, la catégorie [[SpeechAct]], n'est pas présent parmi les trois cadres répertoriés (tableau 3.10).

Schéma de sous-catégorisation	Fréquence	Proportion
[SUJ:SN]	11447	0,497
[SUJ:SN,A-OBJ:SP<à+SN>]	8773	0,381
[SUJ:SN,OBJ:PropSub]	2810	0,122

Tableau 3.10 – Schémas de sous-catégorisation du verbe répondre *dans Lexschem*

En réalité, Syntex n'identifie pas de relations de dépendances entre un discours direct et un verbe de parole comme *répondre*, ce qui augmente considérablement la part de schémas [SN] et [SN Spà]. On ne doit donc pas interpréter ces schémas comme des usages intransitifs (ils sont très rares dans les contes : 0,13), mais comme une absence de détection [Poibeau, 2008 : 65]. On doit donc très certainement intervertir les proportions entre les cadres intransitifs et les cadres complétifs (réalisé en [PropSub] dans le troisième schéma) : la présence d'un message est beaucoup plus fréquente. Nous verrons que cette situation est similaire dans les autres ressources disponibles : le discours direct n'est généralement pas considéré comme un argument syntaxique [Bonami & Godard, 2008]

Pour sa part, DicoValence[19][Mertens, 2010], lexique de cadres valenciels établi manuellement, reporte quatre schémas syntaxiques (tableau 3.11) pour le verbe *répondre*.

Identifiant	Frame	
73200	subj:pron|n:[hum]	?obj:pron|n|compl|inf|de_inf:[abs,mood:ind], ?objà:pron|n:[hum]
73210	subj:pron|n:[hum]	objà:pron|n:[nhum,abs]
73220	subj:pron|n:[hum]	?objà:pron|n:[nhum,abs], ?objp<par>:pron|n:[?nhum,abs]
73230	subj:pron|n:[hum]	objde:pron|n|compl|de_inf:[hum,abs,mood:ind]

Tableau 3.11 – Frames du verbe répondre dans le lexique DicoValence

Le premier « Frame », ou schéma, référence un plus grand nombre de réalisations syntaxiques pour le complément objet du verbe : pronom (*pron*), nom (*n*), complétive (*compl*), infinitive (*inf*), infinitive introduite par de (*de_inf*). Certains cas n'ont pas été détectés dans notre étude, ceci pouvant être imputé à la taille du corpus ou à sa spécificité (bien qu'on trouve cette structure dans des textes de contes ; exemple 112), ou encore à la faible fréquence de ces structures.

(112)La sorcière lui répond de lui faire confiance.[20]

19 http://bach.arts.kuleuven.be/dicovalence/

20 Source : http://merlin.hypnoweb.net/fil-saisons/saison-3/302---poison-mandragore-ii/resume-long-302.158.644/

L'élision du complément objet (le message) est la principale motivation pour la création du schéma 73210, bien qu'il soit optionnel pour le 73200. Ce schéma combine un sujet et un SP en *à*. Pour ce dernier, la contrainte de sélection 'abs' (pour « nom abstrait ») est ajoutée. Elle veut rendre compte de l'alternance que nous avons constatée entre *répondre à une charade* (nom abstrait) et *répondre au chef* (nom humain). Ce complément figure également dans le schéma 73220 (*répondre par la vengeance* ; 107), également illustré en (113) lorsque l'objet indirect désigne le destinataire.

(113)Je lui faisais une révérence en l'appelant citoyenne Michèle, et lui me *répondait* par une révérence encore plus profonde en m'appelant citoyenne Marie.

Le problème est qu'en contexte, un patron comme /[SUJ] verbe [SPà]/ pourrait correspondre à trois schémas différents s'il s'agit d'un humain et à deux schémas s'il s'agit d'un nom abstrait. Comment identifier le bon schéma ? Enfin, le quatrième schéma n'a pas été identifié dans notre corpus ni dans le lexique Lexschem (114) et exprime un sens véritablement différent (référencé dans le Tlfi), « l'engagement à la responsabilité ».

(114)je *réponds* de mes enfants

Une troisième ressource syntaxique peut être consultée à titre comparatif, les tables du Lexique-Grammaire [Gross, 1975] [Tolone, 2011]. Il s'agit d'un lexique distributionnel, où les verbes sont classés (67 tables ou classes) selon les contraintes syntaxiques du verbe et de ses arguments. Un verbe peut appartenir à plusieurs classes s'il possède différents sens. Chaque sens est associé à une « construction de base » qui comporte exhaustivement tous les arguments possibles. Comme DicoValence, le lexique-grammaire utilise des contraintes sémantiques génériques (humain, non humain) mais il identifie six tables ou classes pour le verbe répondre (tableau 3.12).

Table	Construction de base	Contraintes de réalisation (simplifiées)	Exemples
31R	N0 V	N0 n'est pas un être humain	*le téléphone ne répond pas*
7	(N0 V à N1)	N1 peut être un constituant de type proposition	*Cette loi répond à ce que les gens sont mécontents*
33	(N0 V à N1)	N1 ne peut être un constituant de type proposition	*Cet enfant répond à sa mère*
9	(N0 V N1 à N2)	N1 peut être un constituant de type proposition	*Je vous réponds que ça ne se passera pas ainsi*
15	(N0 V de N1 Prép N2)	N1 peut être un constituant de type proposition	*Max répond (devant + auprès de) Luc de ce que tout soit en ordre*
35RR	(N0 V Prép N1 Prép N2)	N1 ne peut être un constituant de type proposition N2 ne peut être un constituant de type proposition	*Max a répondu à cette attaque par le mépris*

Tableau 3.12 – Tables du lexique-grammaire du verbe répondre

Le principal ajout par rapport à DicoValence est la construction intransitive qui requiert que le sujet soit inanimé (table 31R) : pourrait-il s'agir d'un autre nom que *téléphone* ou le contexte téléphonique joue-t-il un rôle déterminant dans l'existence de cette table ? Le formalisme du lexique-grammaire semble requérir également que les constructions où le syntagme régi d'une préposition soit une proposition introduite par *ce que*, soient distinguées comme construction à part entière (table 7). On note néanmoins que parmi les exemples proposés pour la table 33 (équivalent à *X répond à Y*), différents sens du verbe sont mêlés (exemples 115 à 118), parmi lesquels certains sont nouveaux, notamment les deux derniers.

(115)Max répond au téléphone

(116)Cet enfant répond à sa mère
(117)Cet organisme répond au stimulus
(118)Cet objet répond à la description

En (117), le sens du verbe est plutôt de l'ordre de la réaction (sens 3a du Tlfi) alors qu'en (118), il est proche de la notion de conformité (sens 3b-c du Tlfi).
Il ressort de cette courte étude que les patrons, cadres ou structures syntaxiques permettent d'organiser les arguments d'un verbe en fonction des rôles syntaxiques qu'ils remplissent. Néanmoins, nous avons également observé que les critères syntaxiques (position et réalisation syntaxiques) ne suffisent pas à distinguer les différents sens du verbe, ni ceux des arguments. Un des problèmes majeurs est celui de l'ellipse en contexte d'un argument.
Nuançons tout de même que certains lexiques syntaxiques s'appuient sur des restrictions sémantiques, bien que très génériques, pour distinguer des constructions syntaxiques (humain, abstrait). Ce constat invite à explorer le palier sémantique pour identifier des critères systématisables. Deux types d'analyse sémantique sont alors envisageables :

- établir une caractérisation sémantique fine de la nature des arguments d'un patron
- établir une caractérisation sémantique des rôles que remplissent les arguments vis-à-vis de chaque verbe
- mettre à jour des propriétés sémantiques pertinentes partagées par les arguments

Nous explorerons ces pistes dans le chapitre suivant.

4. Patron sémantique en corpus

Afin de parvenir à une description plus fine des constructions verbales, une des alternatives est d'attribuer des catégories sémantiques aux positions argumentales d'un patron donné. Cette perspective s'inscrit dans une tradition linguistique de la description de la structure prédicative [Katz & Fodor, 1963] [Chomsky, 1965] : chaque verbe est associé à une structure prédicative (ou plusieurs), qui représente sous forme de restrictions de sélection la nature sémantique des éléments pouvant apparaître dans une position syntaxique donnée. Dans le cadre de CPA (cf. *infra* 2.4.2), P. W. Hanks propose (dans la continuité des travaux de Pustejovsky ; [Pustejovsky, 1998]) d'utiliser les catégories ontologiques pour exprimer ces restrictions.

Dans ce chapitre, nous décrirons en détail les problèmes posés par la construction/extraction de tels patrons sémantiques à partir de corpus (4.1). Les difficultés rencontrées dans son application nous amènerons à nous interroger sur l'impact du genre dans l'extraction des relations sémantiques (4.2). Nous proposerons trois pistes pour mieux saisir les spécificités sémantiques du corpus et dans la perspective d'adapter des ressources à des corpus spécifiques (4.3).

4.1. Patrons sémantiques ontologiques

4.1.1. Annotation référentielle

La méthode la plus élémentaire pour identifier les contraintes sémantiques susceptibles de s'exercer sur les arguments d'un patron donné, consiste à lister les unités lexicales apparaissant dans une position syntaxique donnée et d'y attribuer la catégorie sémantique qui les satisfasse tous. La principale difficulté à laquelle on est confronté en analyse de corpus, particulièrement pour le français, est la quantité non négligeable de phénomènes de pronominalisation. Le tableau (4.1) liste par exemple les formes des arguments du verbe *dire* (f=731) apparaissant le plus fréquemment en position sujet et objet indirect.

***Sujets du verbe* dire**			***Objets indirects du verbe* dire**		
Forme	**Fréq.**	**Prop.**	**Forme**	**Fréq.**	**Prop.**
il	85	12%	*lui*	170	23%
elle	58	8%	*me*	18	2%
je	32	4%	*te*	10	1%
tu	22	3%	*vous*	9	1%
Zazounette	18	2%	*moi*	5	1%
qui	15	2%	*nous*	5	1%

Tableau 4.1 – Formes des sujets et objets indirects du verbe dire

On constate que les sujets et objets indirects les plus fréquents sont des pronoms (*Zazounette*[21] est un prénom) et qu'ils représentent près de 30% pour les sujets et les objets indirects. Deux possibilités se présentent alors :

- écarter de l'analyse les occurrences pronominalisées, en faisant l'hypothèse que les contraintes que l'on aura identifiées en analysant les unités lexicales « pleines », s'appliqueront également aux pronoms ;
- résoudre la référence en annotant chaque pronom par le nom commun correspondant pour effectuer une analyse sémantique exhaustive des données.

Pour évaluer l'intérêt des contraintes sémantiques, nous avons préféré la seconde solution.

La résolution référentielle consiste en général à identifier les syntagmes co-référentiels, le plus souvent un nom propre et des pronoms personnels identifiant le(s) même(s) individu(s). Ce qui nous intéresse ici est, non pas la co-référence, mais la catégorie ontologique des référents de ces expressions, comme [[Humain]], [[Animal]], etc. Parmi les unités linguistiques annotées, figurent principalement les pronoms anaphoriques et les noms propres ; nous avons également annoté les impératifs, en ajoutant au verbe la catégorie du sujet implicite, et les possessifs (*mon*, *ton*, *son*, etc.).

Nous nous sommes appuyé sur le texte pour choisir la catégorie à annoter. Il est en effet possible, dans la majeure partie des cas, de retrouver une catégorie du référent dans le texte ; dans ces cas, c'est la catégorie la plus précise (*garçon* par rapport à *homme* ; *caméléon* par rapport à *animal*) qui était sélectionnée. Lorsqu'aucune information n'était disponible dans le texte, nous nous sommes à la fois appuyé sur notre interprétation et sur des catégories génériques déjà identifiées. L'exemple (119) illustre le résultat obtenu par cette annotation.

21 La présence d'un prénom aussi incongru est une des conséquences de la taille et du genre du corpus étudié.

(119)"Vous [homme] avez probablement mangé autant de lapins dans votre [homme] vie que moi [animal] et chassé autant de bêtes que nous [animal] tous ensemble."

Dans d'autres contextes, la référence est opaque. Nous avons utilisé une catégorie pour ces indéfinis ([[ONE]]), comme le pronom *on* dans la construction idiomatique de l'exemple (120).

(120)sortit d'on [ONE] ne sait où...

Ce pronom peut néanmoins avoir un référent clairement identifiable, comme un groupe d'enfants dans l'exemple (121) :

(121)C'est sûrement ici que l'on [enfant] va trouver ce que l'on [enfant] cherche !

Un autre type de problème est le choix de la catégorie pertinente pour des entités qui peuvent appartenir à plusieurs catégories référentielles qui s'opposent. C'est notamment le cas lorsque l'histoire évoque des personnages qui se transforment : un homme en prince, ou un prince en dragon, ou encore un crapaud en homme. Nous avons utilisé la chronologie de l'histoire pour sélectionner la catégorie sémantique : en (122), on apprend que désormais le prince sera de la catégorie *dragon* : les occurrences successives du même personnage seront alors annotées en *dragon*.

(122)Un jour, elle enleva un prince qu' [prince] elle transforma en dragon avec une de ces potions.

Cette annotation nous a permis de remarquer que les prénoms n'étaient pas nécessairement garants d'une stabilité référentielle : le nom propre *Christophe*, par exemple, peut, en fonction du conte considéré, désigner un animal ou un être humain, comme l'illustre le tableau (4.2).

Nom Propre	Référent	Type	Fréq.
Christophe	garcon	Humain	8
Christophe	caméléon	Animal	6
Christophe	enfant	Humain	3
Christophe	crapaud	Animal	1

Tableau 4.2 – Lemmes et types sémantiques du nom propre Christophe

Cet aspect référentiel a un impact sur la construction de patrons sémantiques : les contextes attribués à *Christophe*, en tant que caméléon, sont différents de ceux qui sont associés à *Christophe* en tant que garçon, comme le montrent les exemples (123) à (126).

(123)Christophe [caméléon] était méchant, avec des yeux énormes qui clignaient toujours.
(124)Loin de là, dans un grand château, habitait Christophe [caméléon].
(125)C'est l'histoire de Christophe [garçon] un petit garçon beau et gentil mais très coléreux.
(126)Alors Christophe [garçon] partit dans sa chambre tout seul avec ses doudous pour seule consolation.

Le premier référent est un caméléon *méchant* vivant dans un *grand château* alors que le second est un *petit garçon beau* qui possède des *doudous*.

Le tableau (4.3) indique les catégories référentielles majeures du corpus de contes.

Catégorie référentielle	Fréquence
homme	3907
enfant	3451
fille	2159
femme	1546
garçon	734
lutin	521
ONE	489
animal	448
prince	429
chat	421
bûcheron	412
princesse	411
cheval	354
tigre	335
souris	282
roi	282
extra-terrestre	280
fée	274
sorcière	270
hérisson	238
chien	235
vache	225
escargot	201
loup	199
père noël	195
baleine	175
licorne	161
tortue	148
clown	138

Tableau 4.3 – Catégories référentielles employées dans le corpus de conte

On observe que les référents sont majoritairement humain (*homme*, *enfant*, *fille*, *femme*, etc.), mais une proportion non négligeable de référents sont de type animal (*animal*, *chat*, *cheval*, *tigre*, etc.) ou imaginaires (*lutin*, *extra-terrestre*, *fée*, *sorcière*, etc.). En tout, 24688 occurrences ont été référentiellement annotées, soit 15% du corpus (tableau 4.4).

Identifiant	**Forme**	**Catégorie**	**Lemme**	**Type**
149717	Aussitôt	ADV	aussitôt	
149718	remontée	VER:pper	remonter	
149719	Mélisa	NAM	Mélisa	fille
149720	dit	VER:pres	dire	
149721	à	PRP	à	
149722	Fouad	NAM	Fouad	homme
149723	:	PUN	:	
149724	-	PUN	-	
149725	GUILLEMET	PUN:cit	GUILLEMET	
149726	Je	PRO:PER	je	fille
149727	vous	PRO:PER	vous	homme
149728	remercie	VER:pres	remercier	
149729	de	PRP	de	
149730	m'	PRO:PER	me	fille
149731	avoir	VER:infi	avoir	
149732	sauvée	VER:pper	sauver	
149733	.	SENT	.	

Tableau 4.4 – Extrait du corpus de contes

4.1.2. Ontologie et généricité

Les types sémantiques correspondent à des catégories ontologiques génériques, c'est-à-dire que les référents sont répartis en genres et espèces ontologiques (en catégories d'êtres ; cf. *infra* chapitre 1), selon une classification proche des taxonomies employées en sciences naturelles. Si la nature d'une telle ontologie peut être discutée (le terme est aujourd'hui plus élargi et appliqué à de nombreux domaines), on peut faire l'hypothèse que son usage peut s'avérer fructueux dans le cadre d'une tâche de construction de patron sémantique, comme le défend Hanks [Hanks, 2008] : ces concepts seraient des catégories communes de référence. Sachant qu'elle est aussi employée dans les lexiques syntaxiques (3.3.1.) et informellement dans les dictionnaires, la question est de savoir si ce type de propriété sémantico-référentielle constitue une contrainte avérée du verbe.

Nous nous sommes inspiré de l'ontologie de surface de Hanks (cf. *infra* p70) pour faire correspondre à chaque unité lexicale un type ontologique, en la complétant lorsqu'il y avait lieu. Les types sémantiques les plus fréquents sont présentés dans le tableau (4.5).

Type	Fréq.	Exemples
Humain	8248	*père, homme, enfant*
Animal	3031	*chien, baleine, écureuil*
Imaginaire	1551	*lutin, Père-Noël, sorcière*
Lieu	1334	*forêt, château, vallée*
Objet	330	*jouet, chaussure, poupée*
Élément	327	*eau, feu, vent*

Tableau 4.5 – Fréquence des types sémantiques dans le corpus de contes

Nous pouvons faire trois observations concernant cette étape de catégorisation :

- Les types sémantiques sont extrêmement génériques
- Une unité lexicale donnée peut être catégorisée par des types sémantiques différents selon le contexte (*citrouille* est de type [[Végétal]] ou [[Nourriture]]).
- La forte proportion du type [Imaginaire] est propre à ce corpus de contes, et aura des conséquences sur la construction de patrons.

Dans l'ontologie CPA, figure également le type [[Anything]], qui correspond à l'absence de contraintes sur un argument. Par exemple, ce type est utilisé dans le patron CPA le plus fréquent du verbe *to call*, qui peut être traduit en français par *X s'appelle Y* (tableau 4.6) :

31%	**[[Anything]] be called [NOOBJ] {[N]}**
	The name of [[Anything]] is [N] *[[Anything]] may be an individual, a set of thing, a person, a human group, an idea, or anything*

Tableau 4.6 – Patron CPA le plus fréquemment associé au verbe to call *en anglais*

Dans ce cas, le type [[Anything]] indique que toute entité peut être nommée et qu'aucune restriction sémantique ne s'exerce sur le sujet du verbe. On peut conclure que les unités observées

en position sujet de cette forme verbale étaient très variées au point de ne pouvoir choisir d'autre type sémantique que celui-ci. Ce patron n'exploite donc pas d'information sémantique puisqu'il est équivalent à un patron syntaxique tel que /[SN] be called [Npr]/ (où SN désigne tout type de groupe nominal et Npr, un nom propre). En dehors de ce genre de cas, la majorité des patrons CPA manipulent des types sémantiques.

Prenons à titre illustratif le cas du verbe pronominal *s'approcher*. Cette forme verbale apparaît 72 fois dans notre corpus de contes, 30 fois dans le patron /SUJ V de/ et 28 fois sur le mode intransitif /SUJ V/. Les autres patrons relevés sont dérivés de ces deux patrons principaux en y ajoutant une relation de manière (127) ou de but (128), qui peut être considérée comme facultative.

(127)Le minet s'approche *tout doucement* et se faufile dans le feuillage à pas de souris .
(128)Il s'approcha *pour voir ce qu'il se passait*

Générer les différents patrons possibles de ce verbe permet d'observer la variation des types en fonction des positions syntaxiques. Les tableaux (4.7) et (4.8) reportent les fréquences des types sémantiques apparaissant en fonction de la position syntaxique.

Patron 1 [X] se approcher	
ARG1	**F**
[X]	*36*
HUMAIN	*21*
ANIMAL	*10*
IMAGINAIRE	*2*
ELEMENT	*1*
VÉGÉTAL	*1*
SON	*1*

Tableau 4.7 – Patron 1 du verbe se approcher

Patron 2 [X] se approcher de [Y]			
ARG1	***F***	**ARG2**	***F***
[X]	*36*	**de [Y]**	*36*
HUMAIN	*24*	LIEU	*7*
ANIMAL	*7*	OBJET	*6*
ONE	*2*	HUMAIN	*6*
OBJET	*1*	ANIMAL	*4*
VÉGÉTAL	*1*	ACTIVITÉ	*2*
IMAGINAIRE	*1*	VEHICULE	*2*
		ELEMENT	*2*
		MEUBLE	*1*
		VÉGÉTAL	*1*
		IMAGINAIRE	*1*
		CORPS	*1*
		ombre	*1*
		silhouette	*1*
		fissure	*1*

Tableau 4.8 – Patron 2 du verbe se approcher

Les tableaux montrent que les arguments employés dans ces patrons peuvent être de type extrêmement variable. Concernant le patron */X s'approcher/*, on retrouve essentiellement des noms d'animés. Le type le plus approprié pour les sujets de ce patron semble donc être *Animés*, d'autant que cette alternance ne provoque pas de changement de sens du patron : nous avons donc ajouté ce type à l'ontologie, regroupant les types [[Humain]], [[Animal]] et [[Imaginaire]] (fait référence à des êtres imaginaires uniquement). On trouve trois cas uniques qui ne s'accordent pas avec cette hypothèse, les catégories [[Element]], pour le nom *lune* (129), [[Végétal]] pour *plante* (131) et [[Son]] pour *bruit* (130).

(129)La lune remarqua cette courageuse petite fille qui avançait difficilement dans l'obscurité. Doucement elle *s'approcha* et éclaira son chemin, comme pour l'encourager .
(130)Ils entendent un bruit énorme s'approcher, comme un tremblement de terre.
(131)Alors qu'il poursuivait son chemin, il vit des plantes qui lui semblaient s'approcher.

L'exemple (129) peut être considéré comme une figure de style particulière, la personnification, dans laquelle on attribue des propriétés humaines à des entités de nature ontologique différente, soit une exception. L'exemple (131) cherche à traduire une illusion d'optique, justifiée par l'usage du modal *sembler*, qui permet indirectement dans ce cas à une expression dénommant une entité non animée, d'être sujet de ce verbe. En revanche, il semble que le sens du patron en (130) varie : lorsqu'un son s'approche, cela implique peut-être qu'une entité s'approche en produisant du son, mais cela signifie aussi que le son devient de plus en plus audible. Il est donc possible dans ce cas de distinguer ce patron sémantique, portant le sens de « devenir de plus en plus audible » dans lequel le sujet serait de type *son*.

Hormis les exemples de *lune* et de *plante*, le type [[Animé]] semble le plus approprié pour représenter la majorité des occurrences en position sujet dans le patron *X s'approcher*. On obtient donc les patrons sémantiques infiqués en (132).

(132) *Patron 1 :* [[Animé]] s'approcher
Patron 2 : [[Son]] s'approcher

En ce qui concerne le patron /X s'approcher de Y/, on observe le même type de variation pour la position sujet. En revanche, les noms régis par la préposition *de* varient d'une autre manière. On y retrouve principalement des lieux, des objets et des animés. Pour autant, le sens du patron ne change pas car il s'agit systématiquement du sens « s'approcher d'un point de référence concret ». L'analyse des occurrences montre que la propriété d'animé, pour la position complément, n'est pas essentielle pour ce verbe : les animés ne sont pas considérés en tant qu'ils sont capables de se mouvoir de façon autonome, mais seulement en tant qu'ils possèdent un corps ayant une concrétude, positionné dans un espace, dont on peut s'approcher.

Cette propriété de concrétude, ou de matérialité, est partagée par la majorité des unités que l'on retrouve dans cette position argumentale. Nous avons donc créé le type [[Concret]] regroupant les types sémantiques concernés. Parmi les exceptions, on trouve le nom *travail* (133).

(133)Sa maîtresse l'a trop souvent grondée quand elle osait s'approcher de ses *travaux*.

Dans l'exemple (133), il s'agit d'une chienne qui s'approche des tableaux d'arts de sa maîtresse. Dans l'ontologie, *travail* est catégorisé comme [[Activité]], mais désigne ici le produit de cette activité, un objet concret qui sera de type [[Concret]]. Enfin, les noms *ombre*, *silhouette*, et *fissure* figurant en minuscule dans le tableau 4.8 n'ont pas de type sémantique associé. Les deux premiers sont employés pour désigner des surfaces ou des volumes, qui sont trop distants pour être précisément identifiés (on doit donc s'en approcher) ; leur concrétude n'est appréhendée qu'à travers une médiation visuelle. Le nom *fissure*, comme le nom *trou*, sont théoriquement des espaces vides, ce qui peut remettre en cause la notion de concrétude que nous avons proposée. Un tel raisonnement ne nous semble pas pertinent, car les bords de cet espace sont concrets et permettent de définir leur taille : les fissures sont des espaces concrets. Il est donc possible de proposer le patron en (134) pour rendre compte de toutes ces occurrences :

(134) [[Animé]] s'approcher de [[Concret]]

La majorité des alternances apparentes des types sémantiques en position d'argument vis-à-vis du verbe *s'approcher* ont ainsi pu être résolues en créant deux niveaux d'abstraction appropriés dans l'ontologie : [[Animé]] et [[Concret]].

Le problème majeur que nous rencontrons ici est la définition ontologique des êtres imaginaires. L'imaginaire est généralement défini comme ce qui n'existe pas et qui par conséquent est incorporel. Le Tlfi[22] donne par exemple les définitions en (135) de l'adjectif *imaginiaire* :

(135) *Imaginaire :*
1.A. Créé par l'imagination, qui n'a d'existence que dans l'imagination.
Qui ne peut être associé à une figuration concrète.

On retrouve ce type de définition dans les ontologies qui se veulent généralistes : l'aspect fictionnel est écarté dès le départ. Par exemple, la base de connaissance Cyc[23] hiérarchise les êtres fictionnels comme des choses intangibles. Nous reproduisons en figure (4.1) cette hiérarchie.

thing : *Thing is the "universal collection": the collection which, by definition, contains everything there is.*
>partially intangible thing : *The collection of things that either are wholly intangible (see Intangible) or have at least one intangible (i.e. immaterial) part (see intangibleParts).*
>intangible : *The collection of things that are not physical -- are not made of, or encoded in, matter. .*
>non-spatial thing : *the collection of all things with no spatial extent or location, either in some embedding space or relative to some SpatialThing .*
>fictional thing : *The collection of all objects that are TemporalThings (beings, magical artifacts, spells, etc.) in fictional works but not extant in the world modeled by the KB in which something is asserted to be an instance of this. This collection should have no instances in BaseKB Or other general microtheories, but be restricted to microtheories dealing with the "real world" or some fictional world.*

FIG 4.1 – Catégories supérieures aux êtres imaginaires dans la hiérarchie ontologique de Cyc

La définition des « choses fictionnelles » proposée dans Cyc, invite à considérer leur traitement dans des microthéories spécifiques à chaque univers fictionnel[24]. Elles ne sont pas vraies dans le monde réel. Cependant cette catégorie ne nous enseigne a priori rien sur les référents qu'elle regroupe. Comme nous l'avons vu au chapitre 1 (cf. infra 1.2.3), ces entités fictives ou imaginaires sont appréhendées comme réelles : bien qu'elles n'existent pas dans la réalité intersubjective partagée, et qu'elles ne possèdent pas de concrétude en ce sens, elles en sont dotées en discours par la reconnaissance de fait de leur existence. Ces êtres imaginaires, comme le montre notre étude du verbe s'approcher, sont appréhendés comme des animés, certains pouvant posséder un corps.

22 Le Trésor de la Langue Française, cf. http://atilf.atilf.fr/tlf.htm
23 http://www.cyc.com/
24 Les micro-théories, ou contextes, sont des ensembles de concepts et d'assertions relatives à un domaine de connaissance particulier. Le type de micro-théories concerné par les êtres imaginaires porte le code #$FictionalOrMythologicalContext, qui se définit selon l'existence d'au moins une assertion largement jugée fausse (http://www.cyc.com/cycdoc/vocab/context-vocab.html#FictionalOrMythologicalContext).

Par ailleurs, les êtres imaginaires évoluent dans des contextes communs à la réalité : dira-t-on que l'eau, la mer, Paris, décrits dans les contes ne sont pas les mêmes que dans la réalité ? Le texte (la langue ?) ne semble pas distinguer un être imaginaire d'un être humain comme le font les ontologies. On pourra enfin ajouter que le surnaturel n'est pas nécessairement fictionnel : de nombreuses cultures croient en le surnaturel (Magie, Divinités, etc.) et on attribue à certains individus réels comme les sorciers, les shamans ou les marabouts la faculté d'intervenir entre le monde des humains et celui des esprits : s'agit-il d'êtres humains ou d'êtres imaginaires ? La réalité est-elle cette sphère rationnelle où tout peut s'expliquer ? L'étude des contes soulève donc (au moins) deux questions : existe-t-il des corrélations entre distribution sémantique et catégorisation référentielle ? Comment décrire les êtres imaginaires ?

4.1.3. Alternances de type sémantique et métonymie

CPA ne modélise que faiblement les relations sémantiques qu'entretiennent des types sémantiques s'alternant dans une même position : il les homogénéise sous le chapeau d'une catégorie censée capter la propriété ontologique pertinente qu'ils partagent tous. Or, il n'est pas toujours possible d'identifier précisément la catégorie subsumant la totalité des arguments d'un patron donné (comme nous avons pu le faire en isolant les propriétés de concrétude et d'animé). Pour y répondre, Hanks s'appuie sur le modèle du Lexique Génératif (LG par la suite ; [Pustejovsky, 1998]). Il aborde notamment le mécanisme de coercition [*ibid.* : 111], désignant un mode de composition sémantique particulier qui résulte d'un conflit entre le type attendu par un prédicat et le(s) type(s) observé(s). La définition proposée par Hanks et Ježek est la suivante :

> « Type coercion is an operation of type adjustment that occurs when none of the selectional preferences of a predicator match the type of a noun that it combines with in a particular text. In this case, type coercion is invoked to explain how a mismatching verb-argument combination can be interpreted. » [Hanks & Ježek, 2008 : 395]

Ces derniers donnent un ensemble d'exemples tirés de corpus anglais, dont l'alternance [Location]/ [Event] (*lieu / événement*), observée dans la liste des objets du verbe *to attend* :

> « attend
> Direct Object:
> a. [Event]: meeting, wedding, funeral, mass, game, ball, event, service, premiere
> b. [Location]: clinic, hospital, school, church, chapel
> “About thirty-five close friends and relatives attended the wedding”.
> “For this investigation the patient must attend the clinic in the early morning”.
> “He no longer attends the church”. » [*ibid.* : 394]

Les lieux apparaissant en position objet (*church*, *clinic*) sont alors réinterprétés comme les événements qui s'y produisent. En d'autres termes, c'est le verbe qui impose la lecture spécifique d'une unité possédant un type sémantique conflictuel. Ils proposent d'intégrer cette alternance dans le même patron, le sens du verbe dans ce cas restant inchangé (136) :

(136) [[Human]] attend ([[Event | {Location = Functional}]])

Le patron contient une contrainte supplémentaire concernant le rôle (que nous interprétons comme le télique dans la structure Qualia du Lexique Génératif ; [Pustejovsky et al., 2008 : 199]) : certains noms de lieu, comme *church* possèderaient lexicalement la propriété d'être associés à des

événements. On peut néanmoins s'interroger sur la nécessité d'encoder cette information dans le patron verbal. Parmi les alternances de type, Hanks cite [[Human]]/[[Organization]] utilisé abondamment pour le verbe *to deny* (figure 4.2), et, on peut le supposer pour tous les verbes de parole.

No.	%	Pattern / Implicature
1	26%	**[[Human | Institution]] deny [NO OBJ] {that-CLAUSE}** [[Human | Institution]] says that [CLAUSE] is not true
2	15%	**[[Human | Institution]] deny [[Proposition]]** [[Human | Institution]] says that [[Proposition]] is not true
3	6%	**[[Human | Institution]] deny [NO OBJ] {-ING}** [[Human | Institution]] says that he or she did not do [-ING = [[Action Bad]]]
4	1%	**[[Human]] deny [[Action = Bad]]** [[Human]] says that he or she did not do [[Action = Bad]]
5	5%	**[[Human | Institution]] deny {{responsibility | liability} (for [[Event = Bad]]) | knowledge of [[Event = Bad]] | involvement (in [[Event = Bad]])}** [[Human | Institution]] asserts that [[Human | Institution]] is not responsible or liable for [[Event = Bad]], was not involved in it, or did not know about it
6	1%	**[[Human | Institution]] deny [[Request]]** [[Human | Institution]] says no to [[Request]] and says that no action shall be taken with reference to [[Request]] [[Human | Institution]] is in a position of power or responsibility with regard to {request}
7	23%	**[[Human 1 | Institution 1 Eventuality]] deny [[Human 2 | Institution 2]] [[Permission | Privilege | Opportunity | Resource]]** [[Human 1 | Institution 1]] refuses to give [[Permission | Privilege | Resource | Opportunity]] to [[Human 2 | Institution 2]]
8	<1%	**[[Human]] deny ([[Self]]) [[Anything = Benefit]]** [[Human]] practices self-control and does not allow [[Self]] to have [[Anything = Benefit]]
9	20%	**[[Human | Institution | Proposition]] deny [[Abstract = Salient Fact]]** [[Human | Institution | Proposition]] asserts that [[Abstract = Salient Fact]] does not exist or is irrelevant

FIG 4.2 – Patrons CPA du verbe to deny

On constate dans la figure (4.2) que l'alternance sémantique en position Sujet n'a pas été observée pour tous les patrons : les patrons 4 (qui signifie « un être humain refuse d'admettre avoir commis un acte mal intentionné ») et 8 (qui signifie « un être humain s'empêche d'obtenir quelque chose qui lui serait bénéfique en se contrôlant »), tous deux de faible fréquence, sont combinés avec des objets exerçant des rôles sémantiques spécifiques (respectivement *Bad* et *Benefit*). Ces rôles sémantiques correspondent à des noms d'action dont le trait /mal/ fait partie du sens (*meurtre*, *intimidation*, etc.), mais ils désignent plus généralement l'interprétation que l'on fait de l'objet en contexte. La différence entre les patrons 4 et 3 (dans lequel on constate l'alternance Humain/Organisation) ne tient pas à des critères sémantiques différents pour l'objet (ils ont tous deux le rôle *Bad*) mais syntaxiques (l'un est une forme en *-ING*, l'autre un nom d'action) : l'existence de deux patrons repose sur une corrélation syntaxico-sémantique entre le sujet et l'objet. Lorsqu'on consulte les exemples indexés par chaque patron, on observe effectivement une telle corrélation. Par exemple, on ne trouve que des humains pour le patron 4 comme en (137), mais on trouve les deux pour le patron 3 (exemples 138 et 139).

(137)The man denies seven offences of rape and indecency.
(138)Swithland Motors denies sexually discriminating against the men
(139)Stefano has denied pushing his daughter too hard.

On peut se demander s'il ne serait pas possible de trouver des exemples d'organisation pour le patron 4 (à la place de l'humain en 137), c'est-à-dire si cette contrainte syntaxico-sémantique est une norme de la langue anglaise.

Ce phénomène d'alternance sémantique, que l'on peut qualifier de métonymie de type « Institution pour ses membres » (voir « Organization_for_Location » dans [Markert & Nissim, 2007]), occasionne une redondance d'information manifeste, qui se multipliera par le nombre de verbes concernés. Ne serait-il pas possible de le généraliser ?

Nous n'avons pas constaté un tel phénomène dans le corpus de contes. Les organisations sont

rares et les alternances sémantiques majeures concernent plutôt les humains, les animaux et les êtres imaginaires. Par exemple, près de 30% des sujets du verbe *dire* sont des animaux ou des êtres imaginaires (cf. *supra* p120). Lorsque de telles entités « parlent », il est difficile de trouver un lien métonymique comme pour l'alternance Humain/Organisation, qui peut s'expliquer comme un raccourci : un être humain parle au nom d'une organisation et il n'est pas important de signaler l'identité du porte-parole en question.

Mais si ce type de métonymie n'apparaît pas dans les corpus de contes, peut-être n'apparaît-il préférentiellement que dans des textes d'un genre spécifique ? L'alternance sémantique pourrait constituer un moyen de discrimination des genres de texte. Une étude contrastive de corpus est nécessaire pour répondre à cette question.

La comparaison de corpus est une tâche délicate parce qu'elle suppose de contrôler tous les paramètres susceptibles d'influencer les résultats, de savoir exactement ce que l'on cherche à mesurer et de mesurer si les différences sont statistiquement significatives. Le problème est que les conditions sont rarement réunies et qu'il faudrait échantillonner scrupuleusement les textes [Biber, 1993] pour y parvenir. Nous pouvons néanmoins nous forger une idée en comparant les couples syntaxiques d'un même verbe sur deux corpus de genre différent : en étudiant les noms apparaissant dans une relation syntaxique donnée, on pourra identifier de potentielles alternances sémantiques.

Le SketchEngine permet d'accéder au BNC[25], le corpus de référence du projet CPA (d'autres corpus le complètent). Le BNC (100 millions de mots) a fait l'objet d'une description fine des documents, de l'échantillonage des domaines à l'annotation de bruits de fond dans des transcriptions orales. Le corpus intègre une part non négligeable d'oral bien qu'il soit à 90% composé de textes écrits. Le SketchEngine permet de lancer des requêtes sur des sous-parties de ce corpus à partir du type de texte, du domaine ou encore de la date de publication (figure 4.3).

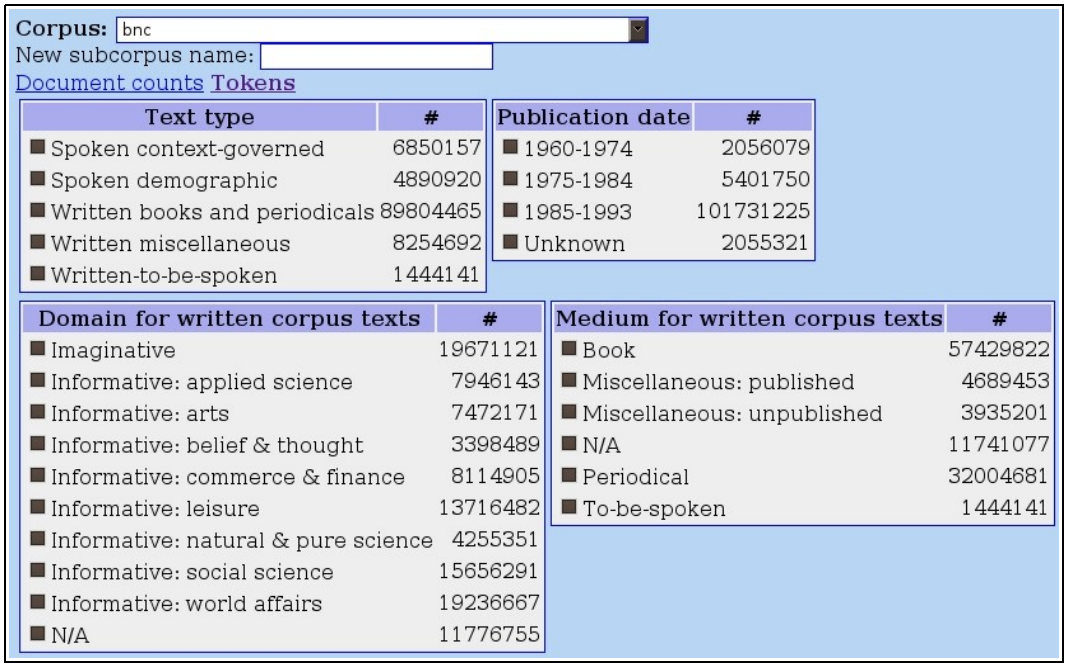


FIG 4.3 – Critères de sélection de documents du corpus BNC par l'interface SketchEngine

25 British National Corpus, cf. http://www.natcorp.ox.ac.uk/

Nous avons cherché à contraster les types de texte relevant de la fiction (domaine *Imaginative*) avec ceux qui sont régulièrement utilisés dans les analyses de corpus ou en TAL, les corpus de presse (domaine *Informative: world affairs*). On constate que leur taille en nombre de mots est relativement équivalente (19,67 millions pour la fiction et 19,23 millions pour la presse ; le nombre de documents étant respectivement de 477 et de 483). L'équivalence est biaisée lorsque l'on s'intéresse au verbe *to deny* : la fréquence absolue de ce verbe est de 2461 (ratio type/occurrence de 0,00012) dans le corpus de presse et de 1004 pour le corpus de fiction (ratio type/occurrence de 0,00005), soit plus du double. Le tableau (4.9) reporte les fréquences obtenues pour les arguments sujets détectés par le SketchEngine du verbe *to deny* dans ces deux corpus triés en fonction de la fréquence de la forme dans ces deux corpus.

Sujet	F1	F2	F3
government	50	0	50
authority	23	0	23
official	21	0	21
spokesman	18	1	19
man	11	0	11
ministry	10	0	10
company	7	0	7
leader	7	0	7
palace	7	0	7
mother	0	6	6
minister	6	0	6
statement	5	1	6
source	6	0	6
soldier	6	0	6
Iraq	6	0	6
Aldington	5	0	5
use	0	4	4
group	4	0	4
army	4	0	4
Smith	4	0	4

Tableau 4.9 – « SketchDiff » des sujets de fréquence globale (F3) supérieure à 3 dans les corpus de presse (F1) et de fiction (F2) pour le verbe to deny

On constate que le corpus de presse (colonne F1) recense un grand nombre de noms d'organisation, comme *government*, *authority*, *ministry*, *company*, etc., qui sont absents du corpus de fiction dans ce contexte. En revanche, ce ne sont que des humains dans le corpus de fiction (*mother, spokesman*), lorsqu'il ne s'agit pas d'erreurs (*use*, *statement*). Il semble donc que l'alternance sémantique Humain/Organisation des sujets du verbe to deny n'existe que dans les corpus de presse. Pour confirmer ce fait, on peut également comparer les verbes avec lesquels un nom d'organisation comme *government*, se combine (Tableau 4.10). Encore une fois, ce qui rend la comparaison difficile est l'écart de fréquence du nom *government* dans les deux corpus (35492 pour le corpus de presse contre 778 pour le corpus de fiction). De plus, on ne peut pas nier que la qualité de la grammaire du SketchEngine n'a aucun impact sur l'identification des sujets (certains couples peuvent être fortuits par exemple).

Verbe	F1	F2	F3
announce	251	2	253
take	248	0	248
make	165	0	165
say	133	0	133
introduce	127	0	127
give	110	0	110
agree	104	3	107
continue	101	0	101
decide	94	0	94
claim	92	0	92
provide	87	0	87
come	79	0	79
set	77	0	77
plan	77	0	77
try	72	3	75
accept	75	0	75
fail	75	0	75
seek	72	0	72
spend	71	0	71

Tableau 4.10 – « SketchDiff » des verbes de fréquence globale (F3) supérieure à 70 dans les corpus de presse (F1) et de fiction (F2) dont le sujet est government

On peut néanmoins observer que ce nom n'est jamais ou que très rarement sujet de verbes de parole (*announce*, *say*, *claim*) ou de verbes qui supposent une intentionalité (*plan*, *decide*), ou encore de verbes d'échange (*take*, *give*, *accept*).

Si ces observations sont validées à plus large échelle, cela implique que des connaissances de l'impact du genre sur les représentations sémantiques peuvent être mises à jour par une analyse de corpus. La seconde conclusion est qu'il n'est peut-être pas souhaitable d'encoder les alternances de types sémantiques dans un lexique sémantique des verbes. La troisième conclusion est que l'obtention de données sémantiques sur la langue suppose un traitement statistique complexe du corpus et de prise en charge des genres. Ceci ouvre la voie à la définition de patrons sémantiques locaux, spécifiques à un genre ou à un domaine.

4.2. Évaluation de l'impact du genre

Nous avons cherché à approfondir l'intérêt des alternances sémantiques dans la comparaison des corpus. Parmi les critères employés dans la classification de texte, on trouve en grande majorité les formes et les catégories morpho-syntaxiques [Biber, 1991] [Santini, 2007]. Nous avons ici voulu analyser l'impact des contraintes sémantiques qu'un verbe exerce sur ses arguments, ce qui, à notre connaissance, n'a pas été évalué dans ce domaine.

4.2.1. Description de l'expérience

La méthode que nous appliquons est la classification ascendante hiérarchique (CAH), ou clustering, sur la base des contextes syntaxiques partagés par les types sémantiques : plus les types sémantiques partageront de contextes, plus ils seront considérés comme étant similaires. Notre hypothèse est que cette méthode permet de mettre à jour la différence d'organisation de types sémantiques et que ce critère constitue une dimension pertinente de variation textuelle.

Pour tester cette hypothèse, nous avons comparé deux corpus de genre différent : l'un, de domaine informatif, un corpus regroupant des articles de presse, l'autre, de domaine fictionnel, le corpus de contes. Les corpus ne varient pas uniquement selon le genre mais également selon le destinataire attendu : adultes pour la presse, enfants pour les contes. Une troisième différence concerne la taille du corpus : le corpus de presse (1 mois du corpus LeMonde, 1 200 000 occurrences) représente approximativement dix fois la taille du corpus de contes. L'équivalence de taille n'est cependant pas une garantie suffisante comme nous l'avons vu pour l'exemple de *to deny*, qui malgré des corpus de taille équivalente, est beaucoup plus fréquent dans l'un que dans l'autre (cf. *infra* 4.1.3). Nous avons cherché à réduire ces différences de la manière suivante :

- Seuls les formes et les types sémantiques communs aux deux corpus ont été analysés.
- Le corpus de contes est annoté manuellement, alors que le corpus de presse est analysé automatiquement par Syntex [Bourigault, 2007]. Ce fait introduit des différences, mais on peut également penser qu'une annotation précise compensera le nombre de couples total.
- Nous n'avons pas pris en compte la fréquence mais la productivité [Bourigault, 2002] : la fréquence indique le nombre de fois qu'un même événement est observé, alors que la productivité indique le nombre d'événements différents observés dans une même configuration. Par exemple, la fréquence des objets directs du verbe *prendre* est de 1021 dans le corpus de presse avec une productivité de 142, alors qu'elle est de 284 dans le corpus de contes avec une productivité de 137.
- La productivité est calculée vis-à-vis des types sémantiques et non vis-à-vis des arguments, ce qui limite les écarts de productivité.

En croisant les deux corpus, on trouve que 120 verbes et 1680 arguments syntaxiques leur sont communs. Ce lexique représente 17% des arguments (en tant que type et non en tant qu'occurrence) pour le corpus de presse et 58% pour les contes. Certaines formes d'arguments sont rares dans un corpus comme dans l'autre. Pour y remédier, nous les avons ensuite associés à un type sémantique dans une ontologie que nous avons adaptée pour les besoins de l'expérience. Cette ontologie comporte 40 types (tableau 4.11).

TYPE SÉMANTIQUE	N	FT	FA	FE	PA	PE	Exemples
LIEU	208	2547	1587	960	189	275	*aéroport, caverne, restaurant*
OBJET	176	1346	445	901	125	252	*chaussure, bague, corde*
HUMAIN	151	9254	1531	7723	223	347	*docteur, homme, enfant*
ÉVÉNEMENT	92	933	712	221	140	91	*randonnée, vol, accident*
CONCEPT_ABSTRAIT	69	1252	1086	166	157	85	*plan, idée, intention*
ANIMAL	66	2996	70	2926	53	314	*cheval, oiseau, chien*
PROPERTY	51	166	121	45	71	27	*magnifique, lent, sage*
EMOTION	47	221	132	89	62	52	*dégoût, sensation,plaisir*
INFORMATION	45	893	701	192	123	60	*histoire, chanson, texte*
TEMPS_PERIODE	44	869	689	180	131	73	*année, hier, début*
CORPS	44	585	232	353	83	147	*bouche, jambe, cheveu*
PERCEPTUAL_OBJECT	40	439	347	92	99	58	*figure, carré, espace*
NOURRITURE	34	282	58	224	35	63	*viande, friandise, gâteau*
VERTU	31	146	90	56	47	37	*fierté, patience, ruse*
VÉGÉTAL	30	432	56	376	35	159	*fleur, sapin, plante*
SOCIAL_ACT	21	526	465	61	83	27	*aide, punition, accord*
VEHICULE	20	208	70	138	39	79	*ambulance, bateau, traineau*
SON	19	304	149	155	56	46	*bruit, cri, fracas*
ACTIVITY	17	282	224	58	74	26	*escalade, jeu, lecture*
SITUATION	16	159	129	30	48	21	*repos, danger, désordre*
SUBSTANCE	16	121	52	69	27	54	*lave, sang, venin*
SUPPORT_D_INFORMATION	15	221	174	47	86	30	*livre, affiche, télévision*
IMAGINAIRE	14	1483	26	1457	19	226	*fée, sorcier, monstre*
METEO	14	166	63	103	41	61	*pluie, orage, neige*
ORGANISATION	13	348	313	35	112	30	*police, entreprise, tribu*
SEPARATEUR_DE_LIEU	12	190	82	108	29	40	*porte, paroi, barrage*
CONTACT	9	133	101	32	43	15	*baiser, fessée, choc*
COULEUR	9	46	30	16	19	11	*rouge, noir, bleu*
RAISONNEMENT	8	135	121	14	47	12	*cause, condition, explication*
COGNITIVE_STATE	8	24	13	11	8	11	*attention, folie, humeur*
ASTRE	6	157	32	125	29	71	*lune, étoile, planète*
EXPRESSION_CORPORELLE	6	77	54	23	29	19	*sourire, démarche, geste*
MÉTAUX_MONNAIE	6	67	55	12	33	11	*or, argent, monnaie*
MAGIC	4	25	10	15	8	9	*magie, sort, miracle*
LIQUID	3	111	31	80	14	56	*alcool, eau, huile*
BREATH	3	30	15	15	6	9	*respiration, souffle, soupir*
ILLNESS	3	11	8	3	8	2	*maladie, épidémie, rhume*
ELEMENT	2	51	19	32	15	22	*air, nuage*
MODE	2	34	29	5	10	4	*accent, ton*
SECRETION	2	4	1	3	1	3	*larme, sueur*
TOTAUX	1376	27274	10123	17151	2457	2935	

Tableau 4.11 – Types sémantiques triés selon le nombre de membres (N), avec FA=Fréq. en Press, PA=Prod. en Presse, FE=Fréq. en Contes, PE=Prod. en Contes et FT= Fréq. totale

Le type Humain n'est pas le plus fréquent, il est devancé par les lieux et les objets. On observe également une répartition différente des types sémantiques : les animaux, objets et imaginaires sont plus fréquents dans le corpus de contes, alors que les concepts abstraits, les événements les informations et les organisations sont plus présents dans le corpus de presse (cette différence se reflète également dans la valeur de productivité). Un des problèmes de cette association forme-type est la gestion de la polysémie : les mots peuvent avoir un sens différent selon le corpus ; de plus certains noms peuvent entrer dans des formules phraséologiques où leur sens est délexicalisé (cf. *infra* 2.3.3) : on ne comparerait dans ce cas pas les mêmes types sémantiques.
La productivité indiquée désigne le nombre de verbes différents liés à un type par une relation syntaxique. Pour calculer la productivité, nous avons combiné le verbe et la relation syntaxique de façon à obtenir une matrice de couples Type sémantique/Contexte. Par exemple, le type [[ILLNESS]] (maladie) a une productivité de 8 dans le corpus de presse et apparaît dans les contextes syntaxiques décrits en (140).

(140)

{apparaître_SUJ, attraper_OBJ, changer_OBJ, comprendre_OBJ, disparaître_SUJ, expliquer_OBJ, passer_SUJ, tuer_SUJ}

Ce type sémantique a une productivité de 2 dans le corpus de contes et apparaît dans les contextes décrits en (141).

(141)

{*attraper_OBJ, prendre_OBJ}*

4.2.2. Calcul des matrices

Avec la méthode d'extraction décrite précédemment, nous obtenons 2457 couples pour le corpus de presse et 2935 pour le corpus de contes. Un tiers de ces couples est commun (978).

Pour réaliser une classification hiérarchique, il faut transformer les données. Le regroupement hiérarchique s'appuie en effet sur une valeur de distance entre les éléments à agréger : ces éléments sont les types et il doivent être comparés. On peut visualiser cette mesure comme la distance entre deux points sur un espace euclidien : le regroupement hiérarchique consistera à associer itérativement les points les plus proches dans cet espace (dont la distance est la plus faible).

Tout d'abord les couples doivent être réunis dans une matrice de taille (m, n) où m est le nombre de lignes (les types, invariable selon le corpus m=40) et n le nombre de colonnes (les contextes syntaxiques ; maximum 347 pour les contes et 223 pour la presse ; cf. *infra* tableau 4.11, p115). La propriété que nous souhaitons évaluer comme indice de (dis)similarité est la productivité : le nombre de contextes syntaxiques différents dans lesquels apparaît un type. La valeur d'une cellule $m_i n_i$ correspond donc à la productivité.

Les mesures de similarité et de distance sont nombreuses. On peut citer par exemple la distance euclidienne d^2 qui s'appuie sur les écarts de deux vecteurs normés (ensemble de ses valeurs, comme la fréquence d'un mot dans chaque document). Pour deux vecteurs A et B, composé de valeurs i (dans notre cas, la productivité d'un contexte), on obtiendrait la formule suivante :

$$d(A,B)=\sqrt{\sum_{i=1}^{n}|A_i-B_i|^2}$$

Dans notre cas, nous souhaitons uniquement savoir si tel type apparaît ou pas dans un contexte syntaxique donné. La valeur que peut prendre la productivité ne nous intéresse pas directement et peut éventuellement introduire un biais dans le calcul de la distance. Il existe un indice qui prend directement en compte dans sa formule cet aspect binaire, l'indice de jaccard. L'indice de jaccard provient de travaux en classification en sciences naturelles (botanique) et consiste à comparer deux objets A et B en fonction de leurs attributs partagées (aussi appelé coefficient de communauté) : par exemple, deux plantes, en fonction d'attributs comme leurs couleurs, leurs milieux, etc. Plus deux objets partageront d'attributs, plus ils seront similaires. Cet indice, ou des dérivés, est notamment employé pour la construction automatique de thesaurus à partir de contextes syntaxiques [Grefenstette, 1994] [Lin, 1998a]; on parle généralement d'analyse distributionnelle en référence aux travaux de Z. Harris : la construction de classes de mots s'appuie sur leurs similarités distributionnelles. L'indice de Jaccard revient à diviser le nombre des attributs communs à deux objets (l'intersection de leurs attributs) par le nombre des attributs de chaque objet

(l'union de leurs contextes). Pour deux types sémantiques A et B, il serait calculé selon la formule suivante :

$$simi_{jacc}(A, B) = \frac{Card(A \cap B)}{Card(A \cup B)}$$

Cet indice est une similarité : plus deux objets partagent de propriétés, plus leur valeur de similarité sera proche de 1. Cette similarité est nulle s'ils n'en partagent aucune. Pour le clustering, nous avons besoin d'une mesure de distance, équivalente à $d_{jacc}=1- simi_{jacc}$. Plutôt que de comparer le nombre de contextes partagés en fonction de l'union des contextes de chaque type sémantique, nous avons normalisé leurs contextes partagés sur le nombre de contextes total pour chaque corpus. La distance entre deux types sémantiques *i j* se calcule donc ainsi :

$$d_{i,j} = 1 - \frac{n_{i,j}}{n}$$

où n est le nombre de contextes total et n_{ij}, le nombre de contextes partagés.

La matrice de dissimilarité est symétrique, de dimensions {m=40, n=40}, où les types figurent en ligne et en colonne, la distance entre le même type étant égale à zéro.

4.2.3. Clustering

La classification ascendante hiérarchique (clustering) consiste à regrouper itérativement une population en partitions (ou clusters) jusqu'à leur épuisement. Les individus de cette population (ou singletons) sont progressivement associés à des clusters et peuvent être visualisés sous forme d'un dendrogramme. Une des hypothèses fondamentales de la classification hiérarchique est la monotonicité : on choisit la meilleure opération d'agrégation à chaque étape, de sorte que les clusters soient de plus en plus dissimilaires.

Il existe plusieurs méthodes d'agrégation (« linkage ») des clusters, dont les plus communes sont l'agrégation unique (« single-linkage »), l'agrégation complète (« complete-linkage ») l'agrégation moyenne (« average-linkage ») et l'agrégation par centroïde (« centroïd-linkage »).

- L'agrégation unique consiste à choisir, parmi l'ensemble des distances entre membres de deux clusters, celle qui est la plus faible (chaque membre d'un cluster est considéré comme représentant de ce cluster).
- L'agrégation complète consiste à agréger en priorité les clusters dont la distance entre les membres les plus distants est la plus faible (le diamètre de ces deux clusters).
- L'agrégation moyenne consiste à s'appuyer sur la moyenne des distances entre pairs de deux membres de deux clusters différents : on agrège en priorité les clusters dont la moyenne est la plus faible.
- L'agrégation par centroïde consiste à choisir la plus faible distance entre les centroïdes (barycentre, centre de gravité, centre d'inertie) de chaque cluster. Le centroïde peut être obtenu en calculant la moyenne des valeurs de dissimilarité des membres d'un cluster (il aura donc une valeur de distance avec chaque membre du cluster).

Ces quatre critères d'agrégation standards sont représentés figure (4.4) .

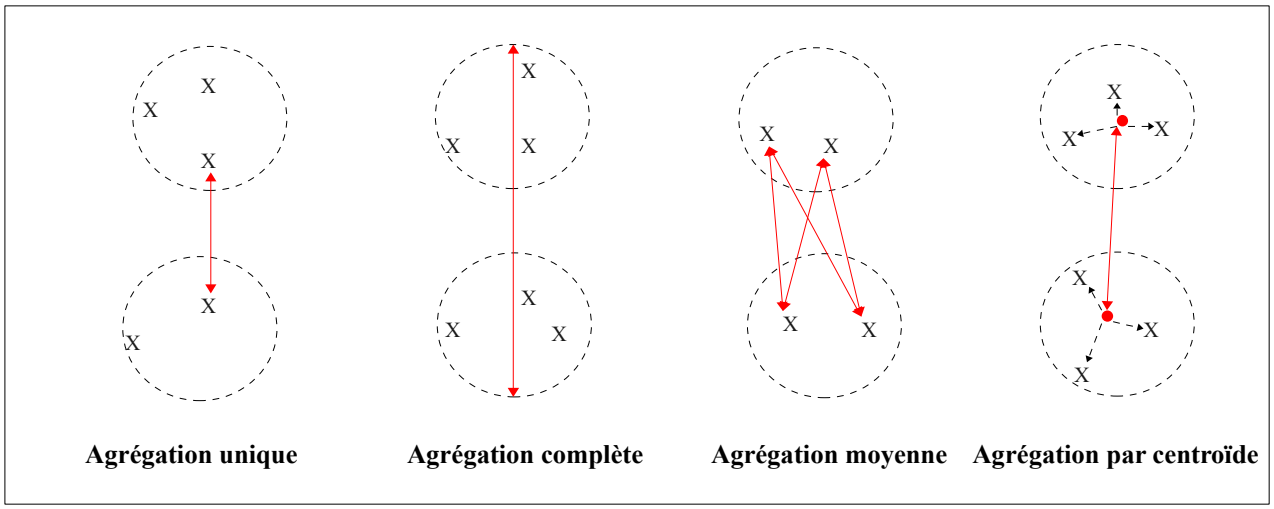


FIG 4.4 – Méthodes d'agrégation standards

La méthode d'agrégation par centroïde a l'avantage de se baser sur une valeur représentative de tous les membres des clusters. Cependant, elle n'est pas monotonique : la distance entre deux centroïdes à une étape t peut être plus faible à une étape t+1 (on obtiendrait un dendrogramme où les lignes se croisent). Les autres méthodes peuvent introduire des biais selon les données. Par exemple, les clusters obtenus par une méthode d'agrégation unique ou complète pourront n'être associés majoritairement que par un membre du cluster. Chaque méthode peut donner des résultats différents.

Nous avons utilisé la méthode de Ward comme critère d'agrégation parce qu'elle tient à la fois compte de la distance entre les centres de gravité (agrégation par centroïde) en minimisant la variance à l'intérieur de chaque cluster. La variance est un indicateur de la dispersion (homogénéité, cohésion) d'un ensemble de valeurs : elle correspond à la moyenne des carrés des écarts à la moyenne et prend donc en compte la distance entre chaque membre et son centre. Une forte variance indiquera une faible homogénéité du cluster. L'indice de Ward associe les clusters en minimisant cette variance, ou inertie. Lorsqu'on regroupe deux classes [i] et [j] en une seule, on peut montrer que la diminution de l'inertie interclasse et donc l'augmentation de l'inertie intraclasse (la somme des deux étant l'inertie totale du système, donc constante) est égale :

$$\Delta_{i,j} = \frac{m_i . m_j}{m_i + m_j} . d^2_{i,j}$$

La méthode de Ward consiste à choisir à chaque étape le regroupement de classes qui minimise l'augmentation de l'inertie intraclasse.
Elle permet d'obtenir un dendrogramme pour chaque corpus que nous allons comparer pour évaluer l'impact des alternances sémantiques.

4.2.4. Résultats

Les dendrogrammes obtenus par cette méthode de clustering pour le corpus de contes et le corpus de presse sont illustrés en (4.5) et (4.6) en fonction de la distance de ward.

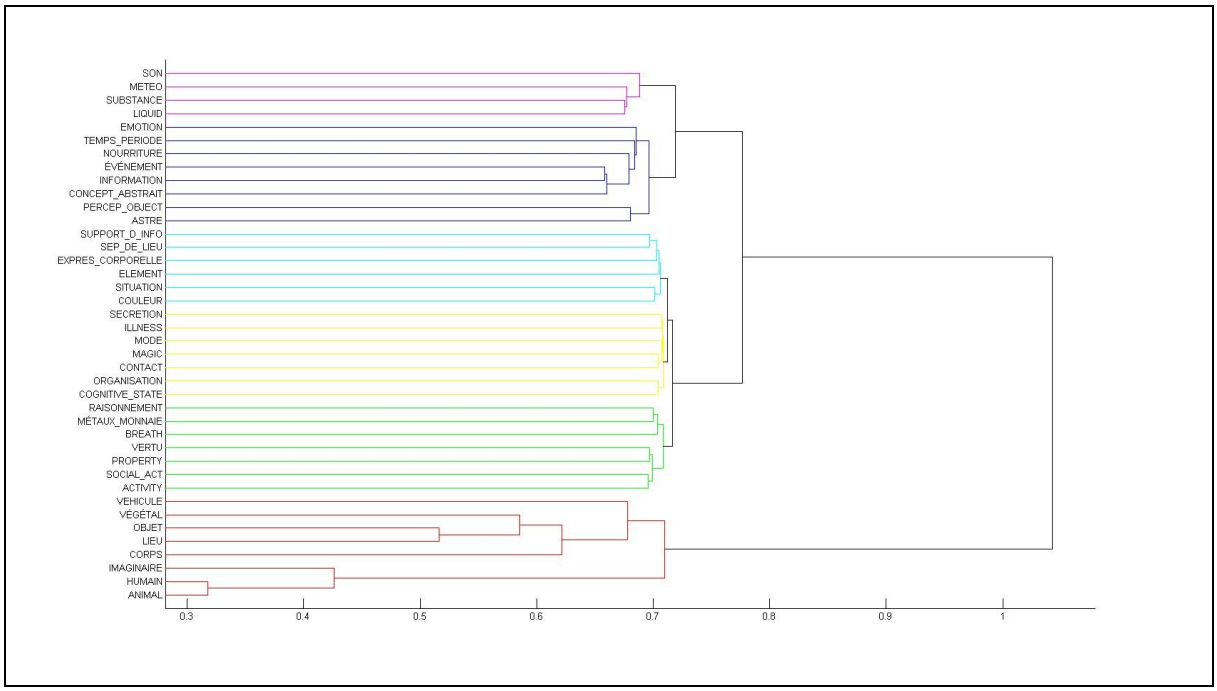


FIG 4.5 – Dendrogramme obtenu sur le corpus de contes (Méthode de Ward)

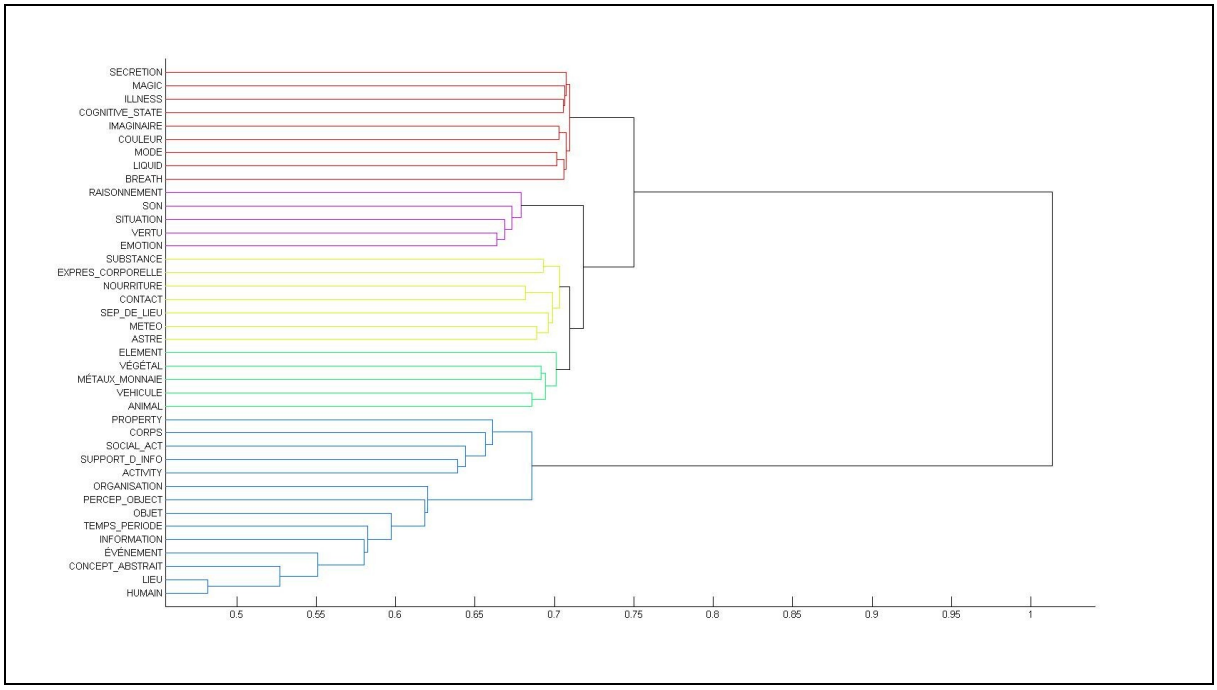


FIG 4.6 – Dendrogramme obtenu sur le corpus de presse (Méthode de Ward)

Les deux dendrogrammes font apparaître que l'agrégation de la plupart des classes s'est effectuée très tard, avec une forte perte d'inertie (distance entre 0,65 et 0,75). On peut interpréter ce phénomène par le fait que la distance choisie ou les contextes ne permettent pas de discriminer suffisamment la totalité des classes, mais on peut également penser que la forte dissimilarité s'explique par un faible nombre de contextes partagés, soit une absence d'alternance sémantique.

Les types agrégés en premier (plus faible distance) sont ceux qui ont la plus forte productivité : Humain (347) et Animal (314) pour les contes et Humain (223) et Lieu (182) pour la presse. On retrouve dans les deux cas le type [[Humain]].

Pour le corpus de contes, les types les plus similaires à [[Humain]] et à [[Animal]] sont [[Imaginaire]] [[Corps]], [[Lieu]], [[Objet]], [[Végétal]] et [[Véhicule]]. En revanche, dans le corpus de presse, les animaux, végétaux et véhicules font partie d'un cluster démarqué qui est fortement dissimilaire au cluster contenant le type [[Humain]].

Les alternances sémantiques ne sont pas du même ordre pour le corpus de presse. Par ordre de dissimilarité, les types [[Humain]]et [[Lieu]] sont associés aux types [[Concept_abstrait]], [[Événement]], [[Information]], [[Temps_période]], [[Objet]], [[Perceptual_Object]] et [[Organisation]]. Mis à part trois de ces types, ce cluster constitue également un groupe cohérent dans le corpus de contes, mais qui est largement dissimilaire au cluster contenant le type [[Humain]] : seuls [[Lieu]] et [[Objet]] font partie du même groupe dans les deux corpus.
Le groupe qui gravite autour du type Humain se distingue également par le fait qu'il est agrégé en dernier avec les autres partitions, et dans les deux corpus.

Cette méthode semble donc confirmer une spécificité de l'organisation sémantique par rapport au genre de texte, tout en révélant des mécanismes similaires (partition du type [[Humain]]).

En analysant les contextes contribuant à leur similarité, nous avons observé deux aspects :

- Ces proximités correspondent régulièrement à des phénomènes de métonymie en position sujet pour le corpus de presse (Humain/Organisation, etc.) et à un phénomène que l'on pourrait qualifier de personnification dans les corpus de conte. Ajoutons que les référents restent des entités concrètes dans le corpus de contes, alors qu'ils sont souvent immatériels dans le corpus de presse.
- Les verbes employés pour établir la classification n'ont pas toujours le même sens : le sens est plus abstrait dans le corpus de presse et plus concret dans le corpus de contes. Par exemple, le sens du verbe *apporter* (ou encore *ajouter*) co-varie très nettement avec le type de ses arguments, type qui coïncide avec le genre de corpus ; il est employé dans le sens de « déplacer un objet en le portant d'un endroit à un autre » dans le corpus enfantin mais la dimension de mouvement concret est ´éludée dans le corpus adulte, comme dans l'exemple (142).

(142)D'après les croyances des peuples de la péninsule , elles sont de bon augure et *apportent* la paix parmi les gens.

Une des façons d'envisager le problème serait de traiter les alternances sémantiques à l'échelle du corpus. À titre d'exemple, les sujets du verbe *dire* illustrent un phénomène d'alternance de type sémantique propre aux contes (tableau 4.12) :

Type	Fréq.	Prop.
Humain	446	62%
Animal	107	15%
Imaginaire	98	13%
Autre	71	10%

Tableau 4.12 – Alternance de type sémantique en position sujet du verbe « dire »

On observe que les humains sont plus souvent sujet du verbe *dire*, mais qu'une part non négligeable des sujets sont des animaux (15%), ou des être imaginaires (13%) et que 10% des sujets sont de type autre (notamment [Objet] et [Végétal]). Ce phénomène peut s'expliquer par une « personnification massive »[26] des personnages de contes, dans lesquels les animaux, par exemple, sont dotés de caractéristiques « habituellement » attribués aux êtres humains, comme la capacité de parler. Il ne conviendrait donc pas dans ce cas de regrouper les entités sous le type [Animé].

Le verbe *dire* n'est pas unique en son genre, cette alternance est également constatée pour les sujets de verbes de parole comme *raconter*, *demander*, *crier*, *appeler*, *parler* et de verbes de cognition comme *savoir*, *comprendre*, *oublier*, *croire*, *connaître*, *décider*.

Pour traiter ces exemples, on peut proposer des opérations de conversion de type spécifiques au corpus. Les êtres imaginaires semblent se comporter comme des humains qui ont en plus, des propriétés surnaturelles (*voler* par exemple). Comme le montre l'exemple (143), les jouets ([Objet]) peuvent aussi être appréhendés comme des humains.

(143) Il écoutait ce que le petit garagiste[jouet] lui [garçon] disait : "Bon, je[jouet] vais essayer de travailler tout seul mais c'est dommage, j'[jouet]aimais bien quand on[garçon, jouet] était tous les deux …"

Le tableau 4.13 résume les principales conversions de type rencontrées dans le corpus de contes (les êtres imaginaires n'y figurent pas).

Type converti	Type de destination
Végétal, Jouet	Animé
Végétal, Animal, Jouet	Humain
Vaisselle, Animal, Végétal	Nourriture
Animé	Son

Tableau 4.13 – Principales conversions de type dans le corpus de contes

Les quatre types de destination des conversions sont *Animé*, *Humain*, *Nourriture et Son*. Les deux premières sont propres aux contes et les deux autres peuvent être rencontrées dans d'autres contextes. Les types Humain et Animé peuvent sembler redondants : on peut considérer que si un type peut être convertible en [[Humain]], il n'est pas nécessaire de spécifier qu'il pourrait également l'être en [[Animé]], parce qu'il en hériterait (un humain est un animé). Néanmoins, différents verbes nécessitent différentes conversions : certains verbes nécessitent la conversion [[Végétal]]→[[Animé]] par exemple, parce qu'ils n'acceptent que des animés (notamment des verbes de mouvement) comme sujet ; d'autres nécessitent la conversion [[Végétal]]→[[Humain]], comme les verbes de parole et les verbes de cognition. De plus, le processus consistant à rendre une entité animée (une « animisation ») est un processus différent de la personnification.

Les conversions vers le type [[Nourriture]] rendent compte du fait que les végétaux comme les animaux peuvent être envisagés dans leur usage nutritif. Les membres de [[Vaisselle]] impliquent que ces objets sont susceptibles de contenir de la nourriture (rapport contenant/contenu). Enfin, la conversion [[Animé]]→[[Son]] traduit la capacité d'animés à produire des sons et à être directement objet de verbes comme *entendre* (144-145).

26 Une personnification est une figure de style. Comme toutes les figures de style, la métonymie par exemple, c'est un procédé exceptionnel qui s'écarte de l'usage ordinaire d'un mot ou d'une structure. Par conséquent, le terme n'est pas véritablement approprié puisque ce phénomène est fréquent.

(144)Mais déjà, dehors, on entend les trolls qui reviennent de leur promenade !
(145)Elle entend le grand sorcier qui ronfle

Ce traitement ne nous paraît néanmoins pas satisfaisant car il a l'inconvénient d'homogénéiser les catégories plutôt que de marquer leurs distinctions. Tous les animaux n'ont pas les mêmes propriétés par exemple et on peut croire que certaines des propriétés pourraient expliquer des combinaisons verbo-nominales.

Nous nous sommes intéressé à l'identification de critères contextuels plus précis pour, d'une part, tenter de contrôler plus systématiquement le sens du verbe, et d'autre part, mieux caractériser les membres d'un type sémantique. Du point de vue de la méthode de classification, cela signifie rechercher des contextes plus précis pour calculer les dissimilarités entre unités sémantiques. D'autres voies de recherche peuvent être explorées en prenant en compte les fréquences dans le calcul de dissimilarité, ou en évaluant l'impact des mesures de distance ou des critères d'agrégation.

4.3. Vers des patrons sémantiques plus spécifiques

Nous avons illustré dans la partie précédente les difficultés rencontrées lorsque l'on s'appuie sur un modèle sémantique dont l'ontologie est générale. Notre hypothèse de recherche est qu'il faut s'orienter vers des modèles plus spécifiques (voir [Kuroda et al., 2006] pour une perspective semblable). Une des alternatives que nous analysons dans cette partie est la prise en compte de catégories sémantiques d'une autre nature que les types : les rôles sémantiques, tels qu'ils sont employés dans la sémantique des cadres (4.3.1). En combinant les informations sur les types et sur les rôles, nous montrons comment on peut parvenir à une description plus fine des unités lexicales (4.3.2). La seconde alternative que nous explorons est l'extraction de propriétés référentielles en utilisant les dépendances entretenues par des unités lexicales dans un même patron (4.3.3).

4.3.1. La sémantique des cadres

La sémantique des cadres constitue une alternative aux modèles sémantiques fondés sur des ontologies, comme CPA ou le Lexique génératif. Les cadres reposent sur le principe suivant : ce sont des systèmes de connaissance, des réseaux de concepts ou catégories auxquels les unités linguistiques donnent accès et sur lesquels elles s'appuient pour leur définition [Fillmore, 1982 : 111]. Par nature, ces systèmes correspondent à des scènes-types de la vie quotidienne (transaction commerciale, par exemple), mais ils peuvent plus largement embrasser tout type de domaine de connaissance présupposé dans la définition d'un concept (voir notamment la définition des domaines chez Langacker [Langacker, 1987]). L'idée de Fillmore est que les unités linguistiques doivent être appréhendées dans le cadre des contextes dans lesquels elles fonctionnent :

> « What I am suggesting here, however, is that linguistics semantics cannot be properly separated from an examination of people's experiences with language in context; and so maybe the two areas of interest are not all that disparate. If, thus, the lexical items we use in our language are essentially items with classifying and describing functions within familiar settings, then there is no critical difference between the two interests. » [Fillmore, 1977 : 72]

Les éléments qui composent les cadres étaient à l'origine très génériques [Fillmore, 2003] : ils correspondent aux cas ou rôles, non plus syntaxiques (sujet, objet, etc.), mais sémantiques, remplis par les arguments d'un verbe. Dans la littérature, ces rôles sémantiques sont également connus sous le terme de « rôles thématiques » ([Dowty, 1991], [Halliday, 1994], [Jackendoff, 1985] par exemple) et ils correspondent à des notions comme Agent, Instrument, Expérienceur, Patient, Thème, Bénéficiaire, Destinataire, bien que leur liste ne soit pas arrêtée. Ces rôles sont liés au sens des verbes : les verbes de mouvement impliquent les rôles Agent et Thème mais pas Expérienceur, là où les verbes de sensation (*percevoir*, *entendre*, etc.) instancieront un argument pour ce dernier rôle.

Enfin, différents niveaux de granularité peuvent être adoptés pour décrire une même situation : plus une situation sera générique, plus le nombre ainsi que la nature des rôles sera réduit. R. Van Valin [Van Valin Jr., 1999] identifie trois niveaux de granularité correspondant aux définitions majeures qui ont pu être proposées dans la littérature, reproduits en figure (4.7).

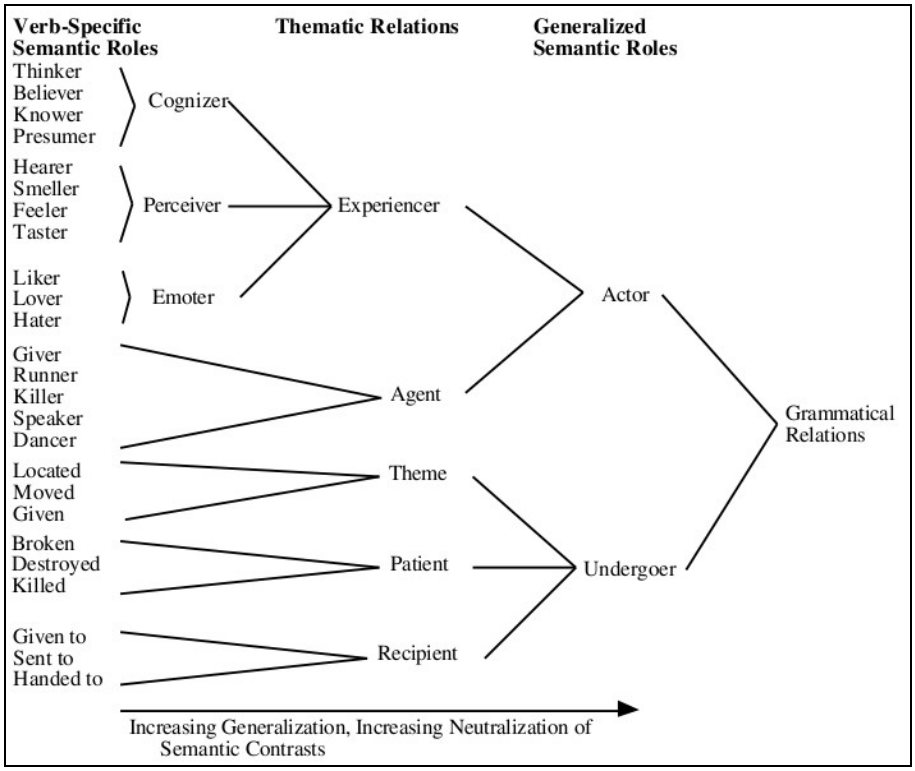


FIG 4.7 – Granularité des rôles sémantiques [*ibid.* : 375]

Néanmoins, ces catégories sémantiques ne sont pas « purement » fonctionnelles car elles reposent en fait sur des présuppositions sur la nature référentielle des éléments qui peuvent remplir ces rôles. En effet, les verbes dénotent des situations référentielles dans lesquelles ne participent que des référents d'un certain type. Le déplacement d'un Thème ne peut s'effectuer que si le Thème est de type Concret, la sensation d'un Stimulus ne peut être ressentie que par une catégorie de référents Animés possédant cette capacité. Fillmore le reconnaît, en qualifiant le fait qu'un agent soit un animé, de « redondance » entre le cas et les traits lexicaux [Fillmore, 2003 : 43]. C'est donc parce que le sens du verbe dénote une situation référentielle particulière que le type sémantique des participants qui y sont impliqués est contraint.

On trouve dans FrameNet[27] les types [[Partie du Corps]], [[Temps]], [[Lieu]] ou [[Événement]] à côté de rôles sémantiques comme {{Agent}} ou {{Theme}}, ainsi que des catégories sémantiques spécifiques au cadre, comme //Crime// ou //Seller//. Le nom commun pour désigner la variété de ces catégories est simplement « élément de cadre » (« frame element »). L'exemple de la transaction commerciale de Fillmore permet de comprendre que plusieurs verbes comme *acheter*, *vendre* ou *payer*, peuvent évoquer un même cadre sémantique dans lequel sont identifiés l'//Acheteur//, le //Vendeur//, le //Bien// échangé, et le //Prix//. Chaque verbe apporte son propre éclairage (ou perspective) sur la scène, mais il relèvent tous d'un cadre plus global, le scénario d'une transaction commerciale.

Notre première tentative d'usage de cette ressource a consisté à identifier les cadres correspondant aux patrons syntaxiques des verbes, et plus spécifiquement chaque élément du cadre à un argument syntaxique. La procédure employée pour chacune des occurrences de chacun de ces verbes est la suivante :

27 Nom de la base de connaissance créée sur ce modèle, cf. https://framenet.icsi.berkeley.edu/fndrupal/

1. identification du sens du verbe et identification du frame correspondant
2. identification des éléments de cadre dans chaque patron

1. Concernant la première opération, la difficulté majeure consiste à choisir parmi les frames disponibles, celui qui restitue le mieux la sémantique propositionnelle. Ce problème est particulièrement saillant dans FrameNet, à cause de la structure en héritage des frames : les frames héritent les uns des autres les rôles et une partie de la sémantique. Parfois la distinction de sens entre deux cadres peut ne pas paraître pertinente. Nous avons reproduit les définitions des cadres //SIMPLE_NAMING// (*SN*) et //BEING_NAMED// (*BN*) pour l'illustrer (146).

(146)

//SIMPLE_NAMING// : *[Entity] [Speaker] [Term]* [Motivation]

« A Speaker **conventionally** uses a particular [*Term*] to **refer to** an [Entity]. »

« This frame concerns **entities conventionally** being **referred to** by particular *names*. »

//BEING_NAMED// : *[Entity] [Speaker] [Name]* [Name_source] [Type]

La différence majeure entre ces deux frames est la substitution de l'élément //Name// du cadre //BN// à l'élément //Term// du cadre //SN//. Il semble que les lexicographes y voient une différence qui s'appuie sur une distinction nom/terme, où « nom » désigne le nom d'une personne, comme le montrent les définitions de l'élément //Entity// dans chaque cadre (147) :

(147) BN : The Entity is the **person** the speaker intends to refer to by using a particular name.
SN : The type of Entity that is conventionally referred to by a Term.

L'analyse des exemples annotés montre à quel point la distinction est délicate : le cadre //BN// est appliqué à d'autres référents que des personnes (148).

(148)//BN// : The device, CALLED a Mubtakker -- Arabic for ``invention'' -- has been at the center of a media firestorm since it was written about in an excerpt from a book extract published by Time magazine last week.
(149)//SN// : The instrument (the telescreen, it was CALLED) could be dimmed, but there was no way of shutting it off completely.

Ce type de phénomène peut nuire à la cohérence globale de la ressource et par conséquent à son utilisation par un système de TAL. D. Gildea et D. Jurafsky [Gildea & Jurafsky, 2002] améliorent d'ailleurs leurs résultats d'annotation automatique de rôles sémantiques en fusionnant les rôles de cadres différents. La similarité des cadres doit donc être étudiée avant l'annotation.

2. La principale difficulté dans la seconde opération est qu'il n'existe pas de correspondance stricte entre cadre sémantique et patron syntaxique. Ainsi, un cadre peut correspondre à plusieurs patrons syntaxiques et un patron syntaxique peut correspondre à plusieurs cadres.

Prenons l'exemple du verbe *chercher* (*f*=349). Dans le corpus de contes, nous avons identifié 176 occurrences de ce verbe dans le sens du cadre //SEEKING//, défini en (150) (signifiant « chercher quelque chose dans un endroit particulier »).

(150) *Définition :*
A [Cognizer_agent] attempts to find some [Sought_entity] by examining some [Ground]. The success or failure of this activity (the [Outcome]) may be indicated. NB: This frame should be compared to the Scrutiny frame, in which the primary focus is on the Ground.

Nous avons identifié ce sens dans pas moins de 23 patrons syntaxiques différents, selon la présence/absence des éléments et selon la nature de leurs réalisations syntaxiques (notamment la variété des prépositions). Les exemples (151) et (152) illustrent chacun de ces deux phénomènes.

(151)Il chercha en Afrique sans rien trouver
(152)Le prince alla chercher le double de la clé dans sa chambre

L'élément //Ground//, qui désigne l'espace d'attention, est réalisé par un syntagme prépositionnel introduit par *en* (151) ou par *dans* (152) et l'entité recherchée (//Sought Entity//) est absente en (152).

Inversement, un même patron syntaxique, comme /[SUJ] chercher [OBJ]/, correspond à 6 cadres différents, ce qui témoigne de la polysémie de ce verbe. Hormis le cadre //Seeking//, on trouve le cadre //Bringing// (dans le sens de « apporter quelque chose »), défini en (153) et illustré en (154).

(153)*Définition :*
This frame concerns the movement of a Theme and an Agent and/or Carrier. The Agent, a person or other sentient entity, controls the shared Path by moving the Theme during the motion.

(154)Va me chercher le diamant bleu, c'est le seul diamant qui manque dans ma collection.

Ce cadre est notamment déclenché par la présence d'un objet indirect qui désigne le bénéficiaire (le pronom *me* en 154).

Dans notre expérience, nous avons parfois été tentés de fusionner les cadres pour en créer de nouveaux, mais pas pour des raisons de similarité de cadre. Par exemple, il existe des cadres pour *chercher* au sens que nous venons de voir (//Seeking//), des cadres pour le recrutement (//Hiring//) ou l'*emploi* (//Employment//), mais on ne sait comment modéliser la *recherche d'un emploi* exprimée en (155).

(155)Après tant d'aventures il était temps de retourner à la maison et de chercher un emploi !

Il est ici question de la composition d'unités lexicales appartenant a priori à deux cadres différents, l'un plus général (//Seeking//), l'autre plus spécifique (//Employment//). La définition du cadre //Employment// (156 ; nous soulignons) justifie son existence mais il ne fait pas partie de la ressource FrameNet.

(156) An Employee and Employer enter into an employment relation, wherein the Employee remains employed for some Duration of time, and finally the relationship ends either by the Employee leaving the job or the Employer letting go (or firing) the Employee. To each of these events there are concomitants, such as agreeing to/signing a contract for entering employment, compensation and performance of a service for the employment period itself, and severance for the dissolution of the relationship. **There are several other events involved, including preparatory actions on the part of the Employer (posting the Position), the prospective Employee's part (looking for a job), or both (job interviews).** In addition, there is the possibility of change in the relationship of Employer and Employee during the employment period, such as a change in Position (e.g. promotion, demotion) and a change in Compensation (e.g. raise, paycut).

Parmi les cadres concernés par l'employé (//Employee's_scenario//), on trouve //Get_a_job//, //Being_employed// et //Quitting//, auquel il semble donc pertinent d'ajouter //Employment_Seeking//. Étant donné que la notion de recherche peut également être appliquée à l'employeur, nous avons également créé un cadre //Employee_Seeking//, pour décrire les exemples (157) et (158).

(157)La Cour cherche des maquilleuses et des couturières
(158)La femme qui la reçut lui dit qu'elle n'aurait rien ainsi gratuitement, mais que précisément elle cherchait une souillon pour s'occuper des dindons, des moutons et des cochons

Contraster l'exemple (155) et (157-158) permet de proposer le type sémantique de l'objet comme élément désambiguïsant : lorsque l'objet est de type [[Humain]] (*maquilleuses*, *couturière*, *souillon*), le cadre est //Employee_Seeking//, le rôle du premier argument étant alors //Employer//. Lorsqu'il s'agit d'une unité lexicale évoquant une activité économique (*emploi*), le rôle du premier argument est //Employee// et le cadre, //Employment_Seeking//. Cependant, le type sémantique n'est pas suffisant pour le distinguer du cadre //Seeking//, dans lequel le second argument (//Sought_Entity//) peut également être de type [[Humain]], comme en (159).

(159)Maintenant, il faut que vous alliez chercher des grandes personnes car je n'arrive pas à m'extraire tout seul de ce buisson.

Pour limiter l'ambiguïté, on peut affiner la catégorie de l'objet du cadre //Employee_Seeking// : la *fonction professionnelle* semble être une propriété partagée des trois occurrences (*maquilleuse*, *couturière*, *souillon*). Le syntagme prépositionnel en *pour* introduisant une infinitive dont le type est une activité, peut venir conforter cette interprétation si le second argument n'est pas suffisamment précis.

Pour résumer, le verbe *chercher* est employé dans les cadres illustrés en (160).

(160) **Employee_Seeking** : //Employeur//, //Fonction_Professionnelle//, //Travail//
Employment_Seeking : //Humain//, //Emploi//
Seeking : //Cognizer//, //Sought_Entity//, //Ground//
Bringing : //Porteur//, //Thème//, //Bénéficiaire// //But//

Comme nous l'avons remarqué, un cadre peut être instancié dans différentes structures syntaxiques, qui peuvent omettre un élément du cadre ou réaliser un élément de diverses manières. Nous avions déjà noté cette difficulté à propos des patrons CPA. Dans ce cas, une description fine de la sémantique des éléments a permis de distinguer les différents cadres. Cependant, les critères de caractérisation syntaxico-sémantique sont mis à l'épreuve à chaque nouvel exemple. Ainsi, les syntagmes prépositionnels introduits par la préposition *pour* désigneront tantôt le //But// du cadre //Bringing// (161), tantôt le //Travail// du cadre //Employee_Seeking// (162) et ce n'est qu'en ayant préalablement identifié les propriétés sémantiques de l'environnement de ce verbe que l'on peut identifier le cadre.

(161)Les paysans courent chercher leurs chariots pour transporter pierres et terre.
(162)La femme qui la reçut lui dit qu'elle n'aurait rien ainsi gratuitement, mais que précisément elle cherchait une souillon pour s'occuper des dindons, des moutons et des cochons

Cette courte étude a montré la nature des difficultés auxquelles on est confronté lorsque l'on cherche à projeter Framenet sur un corpus. Dans la partie suivante, nous chercherons à évaluer l'intérêt d'une telle représentation sémantique.

4.3.2. Filtrage de type selon les rôles sémantiques

Pour expérimenter l'intérêt des rôles sémantiques, nous avons croisé les informations sur le type et le rôle sémantiques. Les rôles (éléments de cadre) ont été approximés à des contextes syntaxiques verbaux spécifiques. Par exemple, on pourra sous-catégoriser les membres des types sémantiques en fonction de leur capacité à être sujets des verbes *marcher* ou *voler*, et en déduire qu'il ont telle ou telle propriété.

Nous avons sélectionné les positions argumentales les plus fréquemment employées par les membres du type [[Animal]] et regroupé ces positions selon l'élément de cadre approprié. Le tableau 4.14 résume les éléments retenus ainsi que la proportion d'animaux occupant ces positions.

Élement de cadre	Exemples de couples	FRS	FA	Prop.
Locuteur	*dire_SUJ, demander_SUJ, répondre_SUJ*	1470	210	14%
Interlocuteur	*dire_à, appeler_OBJ, répondre_à*	813	153	19%
Cognitif	*savoir_SUJ, croire_SUJ, comprendre_SUJ*	809	116	14%
Percepteur	*apercevoir_SUJ, entendre_SUJ, écouter_SUJ*	543	103	19%
Émetteur	*chanter_SUJ, hurler_SUJ, entendre_OBJ*	402	59	15%
Dormeur	*dormir_SUJ, réveiller_OBJ, s'endormir_SUJ*	350	73	21%
Marcheur	*marcher_SUJ, courir_SUJ*	256	76	30%
Mangeur	*manger_SUJ*	191	61	32%
Senseur	*se sentir_SUJ, aimer_SUJ, apprécier_SUJ*	190	50	26%
Sauveur	*aider_SUJ, sauver_SUJ*	181	33	18%
Aliment	*manger_OBJ*	121	23	19%
Forme de métamorphose	*se changer_en, se transformer_en*	120	24	20%
Protégé	*s'occuper_de, aimer_OBJ*	114	23	20%
Joueur	*jouer_SUJ, se diverter_SUJ, jouer_avec*	93	22	24%
Volant	*voler_SUJ, se poser_SUJ*	90	24	27%
Chassé	*courir_après, se sauver_SUJ, courir_IOBJ*	66	30	45%
Entité métamorphosée	*se changer_SUJ, se transformer_SUJ*	65	9	14%
Buveur	*boire_SUJ*	55	9	16%
Combattant	*se battre_SUJ, se battre_contre*	42	18	43%
Assassiné	*tuer_OBJ*	37	19	51%
Marié	*se marier_SUJ, se marier_avec*	36	8	22%
Meurtrier	*tuer_SUJ*	36	2	5%

Tableau 4.14 – Rôles sémantiques retenus associés aux positions argumentales, fréquence en association avec le type Animal (FA) et tout type confondu (FRS)

Le tableau fait figurer la proportion occupée par les unités de type [Animal] par rapport à la totalité des unités occupant cet élément du cadre dans le corpus de contes. On observe par exemple que les animaux occupent une large part des éléments //Assassiné//, //Combattant// et //Chassé//. En revanche, un animal est plus rarement un //Meurtrier//. Les membres *loup* et *chèvre* sont les plus importants //Combattants//, mais les loups occupent préférentiellement le rôle de //Mangeur// alors que les chèvres sont majoritairement perçues comme //Aliment//. Le membre le plus typique des animaux //Chassé// est le veau. Rappelons tout de même que la taille du corpus ne permet pas de généraliser ces observations.

Ces résultats montrent aussi un élément que nous avions constaté pour les êtres imaginaires : dans les contes, les animaux constituent une proportion non négligeable de //Locuteur//, d'//Interlocuteur// et de //Cognitif//. Le type sémantique rencontré entre donc en conflit avec le type attendu (dans 14% des cas pour le locuteur par exemple) et ce phénomène n'est pas prévu par une théorie qui se base uniquement sur une classification ontologique générale. On trouve également des phénomènes propres aux contes : les grenouilles, souris et canard sont les formes de métamorphoses les plus typiques.

Une des questions est de savoir si certaines propriétés exprimées par ces éléments de cadre sont valables en dehors des contes. Comme nous l'avons vu, les animaux sont associés à des rôles liés à la violence (//Assassiné//, //Chassé//, etc.). Mais nous avons aussi observé que les rôles permettaient de distinguer les animaux //Volant// qui sont *aigle*, *papillon* et *oiseau*. Parmi eux, seul *oiseau* a été identifié comme //Marchant//. Pour évaluer la généralité de ces propriétés, il faudrait répliquer cette expérience sur des corpus différents et plus larges.

4.3.3. Dépendance référentielle

D'autres connaissances peuvent être extraites à partir des relations verbo-nominales. Une des propriétés fondamentales des verbes est de mettre en relation des unités qui sont référentiellement liées [Garde, 1985]. Nous avons précédemment vu que les animaux étaient susceptibles d'être des locuteurs mais pas à qui ces animaux parlaient, ou ce qu'ils mangent par opposition à ce que les êtres humains mangent. Comme nous le proposons, ces collocations peuvent nous permettre d'obtenir des informations sur les référents en question.

La dépendance qui nous intéressera est le lien entre le lieu et le type sémantique : certains patrons syntaxiques de verbes comme *vivre* ou *habiter* mettent en relation une entité et un Lieu. Le tableau (4.15) indique par exemple les lieux privilégiés de certains membres de la catégorie [Animal] (la valeur indiquée près du nom d'animal est le nombre d'occurrences associées à des lieux et le pourcentage désigne la proportion des occurrences du nom de lieux associées à ce nom).

Animal	Lieu
chèvre (19)	montagne (14%), ferme (20%), fosse (33%), terre (1%), maison (1%), cage (14%), marché (12%), salle (3%)
tigre (13)	buisson (31%), jungle (38%), branche (5%), magasin (8%)
escargot (12)	feuille (44%), branche (11%), champ (4%), chêne (6%), chemin (2%), salle (3%)

Tableau 4.15 – Les lieux typiquement associés aux animaux

Pour chaque animal, il est possible de classer les lieux, ce qui, en retour, nous permet d'obtenir des propriétés pertinentes de ces animaux. Par exemple, la distinction majeure entre ces différents lieux est Extérieur/Intérieur, ce qui nous permet d'obtenir la classe des animaux sauvages et celle des animaux domestiques. Les tigres et les escargots sont ainsi des membres de la catégorie //Sauvage//, étant donné que 90% des lieux associés à *tigre* comme *jungle* et *buisson* sont des lieux de type //Extérieur//. Pour ce qui est de l'unité *chèvre*, on constate qu'elle est tantôt domestique (*ferme*, *cage*, *maison* et *salle*), tantôt sauvage (*montagne*, *fosse*, *terre*).
En prenant un autre exemple d'animaux, on pourra catégoriser le lézard comme //Domestique// à partir de son association typique avec des bâtiments (tableau 4.16).

Animal	Lieu
lézard (7)	château (4%), maison (1%), chambre (2%), pièce (5%)

Tableau 4.16 – Lieux typiquement associés à lézard

On peut appliquer la même méthode aux membres du type [[Imaginaire]] (tableau 4.17).

Imaginaire	Lieu
sirène (4)	mer (9%), île (5%), rivière (5%)
fée (7)	palais (20%), nuage (11%), château (1%), ruisseau (4%), puits (5%), salle (3%)
sorcière (6%)	laboratoire (20%), chambre (2%), escalier (6%), forêt (1%), cabane (3%), jungle (7%)

Tableau 4.17 – Lieux typiquement associés avec des êtres imaginaires

On constate que les sirènes sont uniquement associées à des lieux //Aquatiques//. Les fées peuvent également être des êtres aquatiques (*ruisseau*, *puits*) mais également aériens (*nuage*). Ils s'opposent tous deux aux sorcières qui sont caractérisées par des lieux terrestres (*jungle*, *forêt*). On peut également observer une autre dimension //Riche// et //Pauvre// en comparant les fées et les sorcières : les fées sont associées à des bâtiments grandioses (*château*, *palais*), à l'inverse des sorcières qui sont liées à des lieux plus modestes (*cabane*).

La dépendance référentielle est notre dernière proposition pour contribuer à répondre aux problèmes sémantiques que l'on rencontre dans les textes. Nous avons vu à travers l'étude des relations verbo-nominales du corpus de contes qu'ils sont nombreux : polysémie verbale, généricité ontologique et divergences syntaxico-sémantiques. Nous avons cherché à complexifier l'analyse collocationnelle, en prenant progressivement en compte les relations syntaxiques, les types et les rôles sémantiques. Chacun de ces niveaux apporte un éclairage différent sur la structure verbale et leur application au texte a illustré certains des problèmes auxquels sont confrontés les théories ou modèles de sémantique lexicale. Pour tenter de les appréhender, nous sommes parvenu à l'idée qu'une approche locale qui prenne en compte la spécificité des textes pourrait permettre de les adapter. Ces considérations nous seront utiles dans les recherches que nous allons à présent aborder car nous y retrouverons des phénomènes similaires.

4.4. Bilan

Nous avons proposé dans ce volet plusieurs analyses de corpus enrichies par différents niveaux d'information linguistique. Ces travaux, destinés à identifier les connaissances nécessaires à l'adaptation d'un système de compréhension dans le contexte du projet EmotiRob, nous ont permis de mieux cerner les avantages et les inconvénients des méthodes d'extraction de relation sémantique employées. Nous avons choisi de ne pas complexifier ces méthodes par des mesures de score, car notre objectif principal était d'évaluer l'apport de la représentation linguistique du contexte. Néanmoins, comme nous adoptons une approche quantitative, les résultats obtenus ont été accompagnés (et les méthodes, guidées), par les mesures de fréquence de cooccurrence et de proportion. Les mesures de score comme l'Information Mutuelle, le test de Student (« t-test ») ou le logarithme de la vraisemblance (« log-likelihood ») ont été employées dans la littérature pour ordonner des couples de mots susceptibles d'instancier une relation sémantique donnée (parfois appelés des collocations). L'application de telles mesures repose sur un certain nombre de choix préalables, comme la taille de l'espace ou fenêtre dans laquelle sélectionner les couples de mots, la prise en compte de l'ordre ou encore de la distance (voir notamment [Washtell & Markert, 2009] pour des approches uniquement fondées sur la distance), dont les incidences nécessitent d'être évaluées. Travailler sur de larges corpus ne suffit pas à justifier l'usage de telles mesures car la majorité des formes (et donc des couples candidats) est rare, ce qui peut biaiser les calculs [Dunning, 1993]. Enfin, il a été montré (par exemple [Krenn & Evert, 2001]) que l'usage de la fréquence de cooccurrence comme mesure d'ordonnancement permet d'obtenir des performances sur des tâches d'extraction de relations sémantiques qui rivalisent, selon l'application, avec des mesures plus complexes.

Le principal biais que nous avons imposé est de nous être focalisé sur les relations verbales. De nombreux travaux ont expérimenté les relations existant entre d'autres catégories (entre un nom et un adjectif ou entre deux noms au sein d'un syntagme par exemple ; cf. : [Nastase & Szpakowicz, 2003], [Nakov & Hearst, 2006]). Nous avons privilégié le verbe parce qu'il peut permettre de mettre en relation plus d'un syntagme et qu'il constitue un riche cadre d'analyse pour les relations sémantiques. Par ailleurs cela répondait aux besoins de l'application, EmoLogus nécessitant en priorité des connaissances sur ces relations.

L'approche linguistique de modélisation du contexte suppose également d'effectuer des choix. Pour ce qui concerne l'annotation syntaxique, nous avons par exemple choisi d'annoter les segments de discours direct, ce qui nous a permis de faire émerger les différences entre nos patrons syntaxiques et ceux extraits automatiquement de corpus automatiquement annotés comme Lexschem. Nous avons choisi de ne pas établir une frontière entre argument et circonstant (voir par exemple [Fabre & Frérot, 2002]) parce que nous souhaitions être aussi exhaustif que possible sur les relations verbales observées en contexte (quitte à en écarter pour des applications ultérieures).

L'étude des patrons sémantiques nous a permis de montrer, dans un premier temps, que les types ontologiques constituaient des propriétés pertinentes pour la description des relations sémantiques verbales ; en d'autres termes, qu'il existe des corrélations entre les propriétés distributionnelles des arguments d'un patron syntaxique et les propriétés ontologiques des référents désignés. Des études doivent être menées à plus large échelle pour confirmer cet aspect, mais il est nécessaire au préalable de se prononcer sur le problème de la modélisation des alternances sémantiques, qui constitue la principale difficulté pour établir des patrons ontologiques. De notre point de vue, ces alternances ne doivent pas être encodées dans un lexique général car elles relèvent, comme nous avons essayé de le montrer, d'un type de corpus particulier. Les catégories ontologiques sont par conséquent parfois trop génériques et ne suffisent pas à désambiguïser les contextes. Il est ainsi possible que nos observations initiales soient le fruit d'une coïncidence ou

d'une tendance, mais il semble nécessaire, au terme des analyses proposées, de devoir adapter ces ressources génériques en fonction des corpus au moyen d'un modèle qui capte ses spécificités. C'est dans ce sens que nous avons proposé des méthodes pour acquérir des propriétés sémantiques adaptées au corpus.

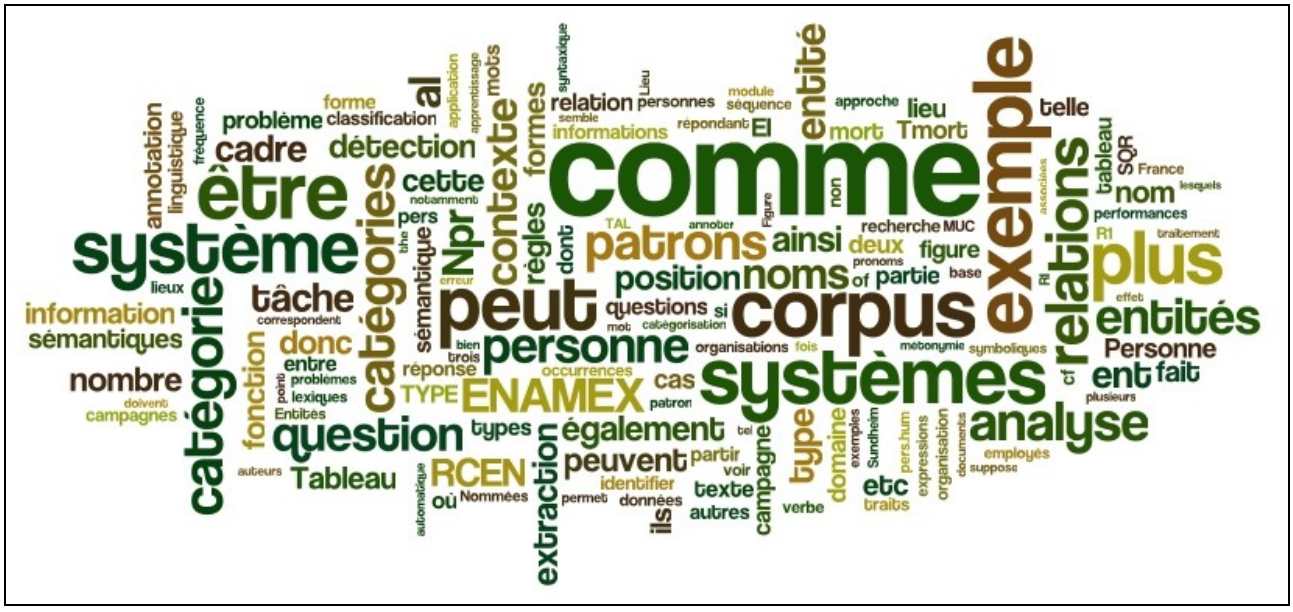
L'analyse des similarités entre types ontologiques selon le genre de texte s'inscrit dans un plus large champ de recherches qui concerne les approches distributionnelles. Les systèmes, qui s'appuient sur les proximités textuelles ou syntaxiques partagées par deux mots, sont généralement évalués sur des ressources lexicales, comme des dictionnaire de synonymes (voir [Lin, 1998b] ou [Bourigault, 2002] pour le français par exemple). Nous avons montré que la distribution pouvait être employée pour caractériser les différences d'alternances sémantiques observées dans les textes. Cette expérience nous a conforté dans l'idée que les types ontologiques ne suffisent pas à représenter les relations sémantiques en corpus et nous a amené à identifier d'autres catégories sémantiques comme les rôles pour modéliser les relations en contexte.

TROISIÈME VOLET

SÉMANTIQUE POUR L'EXTRACTION D'INFORMATION

Le Traitement Automatique des Langues (TAL) est un domaine de recherche au carrefour de la linguistique et de l'informatique. Orienté vers la conception d'applications ciblées (recherche d'information, traduction automatique, résumé automatique, etc.), son matériau d'analyse principal est le texte (supports oral, écrit, etc.). Les programmes informatiques rencontrent inévitablement des problèmes de sémantique. Ces applications peuvent être assistées par une modélisation linguistique du texte, mais elles font aussi émerger des problèmes spécifiques et stimulants d'un point de vue linguistique.

Un des domaines de TAL auxquels la sémantique a pu et peut encore contribuer est l'Extraction d'Information (EI). Le chapitre 5 présentera ce domaine en focalisant sur la détection d'Entités Nommées. Il nous permettra de souligner les liens avec les phénomènes décrits dans le volet précédent et de préciser notre approche. Le chapitre 6 fait état des expériences que nous avons réalisées pour répondre aux problèmes rencontrés dans ce domaine. Nous y développerons notamment les problèmes auquels devra répondre notre système d'Extraction d'Information évalué dans le quatrième volet.



5. Extraction d'information et Entités Nommées

Nous pouvons définir préalablement l'EI comme une tâche de TAL consistant à repérer, catégoriser et associer des informations-cibles contenues dans une source textuelle non-structurée. Les systèmes d'EI sont conçus pour répondre à un besoin informationnel particulier et ne prétendent pas assurer une compréhension totale des textes. Le cadre le plus commun dans lequel ces systèmes sont employés est la recherche d'information précise [Zweigenbaum et al., 2008], qui nécessite une chaîne de traitement complexe. Par exemple, les systèmes de Question-réponse doivent traiter une question et identifier la réponse (ou les réponses) dans une base de documents. C'est également dans ce cadre que nous situerons notre recherche. L'EI est à la fois confrontée à des problèmes de représentation des connaissances et d'ingénierie linguistique. La première partie de ce chapitre introduira ce domaine : nous verrons que les Entités Nommées (EN) y occupent une place centrale. Dans la seconde partie, nous nous intéresserons au problème de classification sémantique des EN, en caractérisant ses rapports avec les ontologies, le discours et le contexte d'application. Pour finir nous décrirons les approches développées en TAL pour la reconnaissance et la classification des EN (RCEN) et préciserons l'approche que nous adoptons.

5.1. Origine et définitions des Entités Nommées

5.1.1. Les cadres de l'extraction d'information

L'extraction d'Information a émergé de besoins d'accès au contenu des documents [Poibeau, 2003]. Comme le rappellent R. Gaizauskas et Y. Wilks [Gaizauskas & Wilks, 1998], l'EI doit être distinguée de la Recherche d'Information (ou RI), par les objectifs et les méthodes qu'elles emploient. La RI est traditionnellement une approche relativement pauvre en connaissance, qui utilise les concepts de la Théorie de l'Information [Shannon, 1948] pour retourner un ensemble de documents pertinents vis-à-vis d'une requête ; le texte est conçu comme un sac de mots constituant ses unités d'indexation. À l'inverse, l'EI est typiquement une application de TAL, qui hérite des recherches menées en Intelligence Artificielle sur les systèmes à base de règles, dits symboliques, riche en connaissances ; l'objectif n'est plus de retourner des documents, mais d'extraire des informations structurées à partir de texte, ce qui suppose une analyse des mots en contexte.

Les domaines d'application de l'EI peuvent être extrêmement divers. Pour ne citer qu'un exemple, le domaine médical a donné lieu à un large corps de travaux, dont les premiers, d'après N. Sager, remontent à 1969 [Sager, 1982]. Les textes de rapports cliniques par exemple sont rédigés dans un style spécifique et les connaissances qui y sont véhiculées, comme les symptômes, les maladies sont propres au domaine (pour des travaux récents, voir [Grouin et al., 2011]). Nous ne nous intéresserons pas particulièrement à ces sous-langages dans cette thèse, mais à des informations qui relèvent de domaine plus généraux et tels qu'on peut les trouver dans des corpus de presse.

Une manière commune de concevoir L'EI est de modéliser le type d'information recherchée sous forme de cadres ou grilles (« templates ») pour lesquels les systèmes remplissent des champs d'information, comme pour une base de données. C'est du moins ainsi qu'ont été proposées les premières campagnes d'évaluation d'EI, MUC (Message Understanding Conference).

L'une des premières tâches d'EI était l'extraction d'informations sur les attaques terroristes dans des corpus de presse[28] (MUC-3 ; [Sundheim, 1991]). Il s'agissait d'identifier le type d'incident (parmi une liste) rapporté dans un article, ainsi qu'un certain nombre d'informations concernant ce cadre. Le cadre était organisé en fonction des rôles majeurs liés à ce cadre (les champs), l'Incident (« Incident »), son Auteur (« Perpetrator »), la Cible Physique (« Physical Target ») et la Cible Humaine (« Human Target »), comme indiqué dans le tableau (5.1).

28 Les corpus employés pour développer et évaluer les systèmes peuvent être considérés comme spécialisés, dans la mesure où ils ont été sélectionnés à partir de requêtes contenant les mots comme *explosion* et un nom parmi 9 pays considérés [Sundheim, 1995 : 6 et 16]. L'auteur considère néanmoins que la moitié des articles (environ 1300) n'était pas pertinente vis-à-vis de la tâche, pour des raisons de polysémie des termes de la requête par exemple.

Champ	Exemple de fiche
MESSAGE: ID	DEV-MUC3-0718 (UNL/USL, NCCOSC, GE)
MESSAGE: TEMPLATE	1
INCIDENT: DATE	- 13 NOV 89
INCIDENT: LOCATION	EL SALVADOR: SAN SALVADOR (DEPARTMENT) : SOYAPANGO (CITY): PRADOS DE VENECIA (NEIGHBORHOOD)
INCIDENT: TYPE	BOMBING
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT ID	-
INCIDENT: INSTRUMENT TYPE	BOMB: "-"
PERP: INCIDENT CATEGORY	STATE-SPONSORED VIOLENCE
PERP: INDIVIDUAL ID	-
PERP: ORGANIZATION ID	"AIR FORCE"
PERP: ORGANIZATION CONFIDENCE	REPORTED AS FACT: "AIR FORCE"
PHYS TGT: ID	"HOUSES"
PHYS TGT: TYPE	CIVILIAN RESIDENCE: "HOUSES"
PHYS TGT: NUMBER	PLURAL: "HOUSES"
PHYS TGT: FOREIGN NATION	-
PHYS TGT: EFFECT OF INCIDENT	DESTROYED: "HOUSES"
PHYS TGT: TOTAL NUMBER	-
HUM TGT: NAME	-
HUM TGT: DESCRIPTION	"INJURED" "PERSONS"
HUM TGT: TYPE	CIVILIAN: "INJURED" CIVILIAN: "PERSONS"
HUM TGT: NUMBER	12: "INJURED" 3: "PERSONS"
HUM TGT: FOREIGN NATION	-
HUM TGT: EFFECT OF INCIDENT	DEATH: "PERSONS" INJURY: "INJURED"
HUM TGT: TOTAL NUMBER	-

Tableau 5.1 – Cadre-scénario de la campagne MUC-3 [*ibid.* : 8]

On peut observer une forte similarité entre ces grilles et les cadres sémantiques (cf. *infra* 4.3), mais les informations y sont plus nombreuses et plus précises. Par exemple, les événements concernés sont classés (« Incident Type ») : les systèmes doivent déterminer s'il s'agit d'un incendie, d'un bombardement, d'un enlèvement, d'un détournement (de véhicules), d'un cambriolage, de grèves ou d'attaques ; ils doivent également déterminer si l'auteur de l'incident est une institution étatique ou pas (« Perp : Incident Category »). L'une des catégories qui va prendre de plus en plus d'importance dans le cadre des campagnes MUC est le nom ou l'identifiant des acteurs de ces événements. Les campagnes MUC qui suivront verront naître une nouvelle tâche complexe et auparavant relativement délaissée par les linguistes comme par la communauté de TAL : le traitement (sémantique) des noms propres *dits* Entités Nommées.

5.1.2. Les entités nommées

C'est lors de la sixième campagne MUC [Grishman & Sundheim, 1995] que des sous-tâches de l'EI ont été distinguées. Au lieu de proposer uniquement une grille de scenario-type à remplir comme cela avait été le cas pour les campagnes précédentes, les organisateurs ont défini trois tâches complémentaires : la tâche d'Entités Nommées, la tâche d'Éléments de Cadre, la tâche de Co-référence. Nous décrirons uniquement la première.

La tâche d'Entité Nommées a émergé d'un besoin d'identifier des sous-tâches d'EI suffisamment indépendantes du domaine considéré pour être réutilisables. Les EN (ou Enamex pour « Entity name expression ») étaient de trois types (Personne, Lieu, Organisation)[29] et la tâche consistait à insérer des balises SGML autour de chaque nom propre relevant d'un des types, comme illustré en (163) (exemple tiré de [*ibid.* : 6]).

29 D'autres entités ont été ajoutées : les moments et dates (TIMEX), et les montants et pourcentages (NUMEX), ce qui fait un total de 7 avec les EN.

(163)Mr. <ENAMEX TYPE="PERSON"> Dooner </ENAMEX> met with <ENAMEX TYPE="PERSON"> Martin Puris </ENAMEX>, president and chief executive officer of <ENAMEX TYPE="ORGANIZATION"> Ammirati & Puris </ENAMEX>, about <ENAMEX TYPE="ORGANIZATION"> McCann </ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE="MONEY"> $400 million </NUMEX>, but nothing has materialized.

Les systèmes étaient évalués en fonction des mêmes métriques employées pour les campagnes MUC précédentes : la précision, le rappel et la F-mesure qui sont calculées comme suit :

$$Précision = \frac{Nombre\ d'éléments\ correctement\ identifiés}{Nombre\ d'éléments\ correctement\ identifiés + Nombre\ d'éléments\ mal\ identifiés}$$

$$Rappel = \frac{Nombre\ d'éléments\ correctement\ identifiés}{Nombre\ d'éléments\ total\ à\ identifier}$$

$$F - mesure = \frac{(1+\alpha).Précision.Rappel}{\alpha.Précision + 1.\text{Rappel}} \quad avec\ \alpha > 0$$

La F-mesure est un indicateur combiné des performances du système. Le paramètre α est choisi en fonction du poids que l'on souhaite donner à la précision ou au rappel ; dans les travaux de TAL, on trouve souvent 2 au numérateur, ce qui équivaut à leur donner le même poids (α=1). Cette mesure donne une meilleure idée que la précision ou le rappel pris indépendamment. En effet, considérer tous les mots d'un texte comme des Entités Nommées permet d'obtenir 100% de rappel mais une précision médiocre ; en revanche, ne considérer que les mots dont on est sûr comme des EN permet d'obtenir une bonne précision mais un rappel médiocre.

En réalité, lors de cette campagne, ces mesures ne donnaient pas les résultats exacts mais combinaient deux dimensions de score : la détection et la classification. La détection concerne la délimitation correcte des frontières d'une EN alors que la classification concerne l'attribution d'une catégorie. Détecter une EN est généralement plus aisé que la classer, et, dans un tel système, le score du nombre d'éléments correctement identifiés augmente même si la détection est erronée (voir [Nadeau & Sekine, 2007]). Ce mode de calcul donne donc une vision plus tronquée de la classification des EN qu'une mesure qui exigerait une détection et une classification exactes.

La majorité des systèmes ayant participé à la campagne MUC-6 ont obtenu des performances supérieures à 90% de F-mesure, et, pour certains, à hauteur de l'annotation humaine [Sundheim, 1995]. Des résultats similaires ont été obtenus pour la campagne suivante, MUC-7.

On trouve dans le guide d'annotation de ces campagnes des indications qui permettent de clarifier plus exactement la nature des expressions linguistiques à annoter ainsi que les principes à adopter pour attribuer les étiquettes. Les EN considérées sont principalement des noms propres ; les titres et rôles d'une personne, par exemple, sont exclus de l'annotation :

> « Titles such as "Mr." and role names such as "President" are *not* considered part of a person name. However, appositives such as "Jr.", "Sr.", and "III" *are* considered part of a person name. » [Chinchor, 1998]

Les textes étaient normalisés [Sundheim, 1995 : 16] comme on peut le voir dans l'exemple (163), à l'inverse des précédentes campagnes pour lesquelles le texte était écrit en lettres capitales, ce qui facilite la détection des EN. Un des systèmes a évalué à 10% la baisse de performance (en termes de F-mesure ; [*ibid.*]) sur une version du texte entièrement capitalisée. Le nombre d'EN

évaluées est de 925, dont 48% d'organisations, 40% de personnes et 12% de lieux pour 30 articles.

L'analyse des erreurs commises par les systèmes lors des sixième et septième éditions a mis en valeur deux difficultés principales. Comme tous les noms propres ne doivent pas être typés, certains, comme les noms propres d'événements ou de produits (journaux) pouvaient constituer une source d'erreur. Le problème de l'ambiguïté du type en contexte est également soulevé, sans que soit pour autant mesurée son importance : les difficultés concernent la détection d'organisation, pouvant être désignée d'après une personne ou d'après un lieu.

5.1.3. Définition et Usage

L'analyse des Entités Nommées dans le cadre des campagnes MUC est critiquable à bien des égards mais il faut reconnaître qu'elle a permis d'exposer au grand jour une problématique linguistique jusque-là largement ignorée : qu'est-ce qu'un nom propre ? La scission entre dictionnaire et encyclopédie est déjà révélatrice d'un doute sur la nature linguistique des Npr, puisqu'ils sont généralement exclus du premier et figurent majoritairement dans le second. Jusqu'à récemment, la théorisation du Npr était une préoccupation très éloignée des linguistes : c'était un problème de Logique, de Référence, donc de Pragmatique, ou alors son traitement était abordé dans le cadre de l'onomastique, discipline qui étudie notamment ses origines étymologiques, ses dimensions dialectales et géographiques. Le Npr, comme nous l'avons vu (cf. *infra* 1.2) est lié, à l'origine, aux théories référentielles du sens. Il a récemment connu, en France particulièrement, de nouvelles descriptions, sous la large bannière de « dénomination » [Gary-Prieur, 1994] [Jonasson, 1994] [Kleiber, 1995] [Noailly, 1995]. Bien que les analyses se concentrent majoritairement sur les noms de personne, ou anthroponymes on trouve aussi des analyses de noms de lieux, de noms collectifs ou encore d'événements [Lecolle et al., 2009]. Pour ce qui concerne la linguistique de corpus, nous n'avons trouvé aucune étude pertinente sur la catégorie des Npr (voir tout de même [Pierini, 2008]).

En TAL, l'analyse des Entités Nommées est essentiellement considérée comme un problème de catégorisation [Ehrmann, 2008 : 2.1] : il s'agit de déterminer si une occurrence (bien souvent un Npr) appartient à l'une des catégories prédéterminées par les conventions d'annotation. Ces catégories sont exclusives : une séquence est ou une personne ou un lieu, mais pas les deux à la fois. Ce principe d'exclusion témoigne d'une rigueur ontologique, ou d'une stabilité référentielle si l'on préfère. Les EN sont des objets du monde réel, des référents, pour lesquels les locuteurs possèdent un « identifiant unique » [Chinchor, 1998], un Npr, qui les distingue d'autres référents pouvant appartenir à la même catégorie. Plutôt que de nous lancer dans un recensement des divers formules de définitions proposées dans la littérature, il nous semble plus approprié de définir ce concept par l'usage pratique qui en est fait comme nous avons commencé à le faire en étudiant les guides d'annotation (pour les définitions, voir [Ehrmann, 2008 : 3.3]).

Les EN traitées dans la campagne MUC devaient être des « identifiants uniques », signifiant par là à la fois l'unicité du référent et l'unicité de l'identifiant assurant une stabilité référentielle vers ce référent. Par exemple, le guide de MUC-7 indique que des lieux spécifiques comme *la Ruhr* doivent être annotés, parce que l'entité est identifiable même lorsque celle-ci est décontextualisée (en dehors de tout texte).

> « Do tag names of sub-national regions when they are associated with specific regions, if they are identifiable even when the name is disassociated from context. » [Chinchor, 1998]

Les formes que peuvent prendre les EN sont conditionnées par cette convention, doivent être invariables et toutes portent une majuscule : les EN à annoter sont des Npr, des initiales et des acronymes (*IBM*).

Ces guides fournissent des principes pour caractériser l'étendue syntagmatique d'une EN. Les choses seraient trop simples si ces identifiants uniques correspondaient à des Npr séparés par des espaces. Une combinaison /prénom + nom/ par exemple comptera pour une entité de type Personne. Si un Npr de personne est suivi de *Jr.* (abréviation de *junior*), ce dernier est inclus, mais s'il est précédé de « désignateurs » comme un titre (*Mr.*, abréviation de *Mister*) ou une fonction (*President*), ces derniers sont exclus. Pour ce qui concerne les lieux et les organisations en revanche, le désignateur peut être inclus. Un point cardinal (*North* en 164) ou un désignateur (*River* en 165) seront inclus dans une EN de Lieu s'ils font partie « intrinsèquement » du nom. Les désignateurs d'organisation comme *Inc.* ou *Co.* feront toujours partie de l'EN (166-167) et déterminent sa frontière, ce qui a pour conséquence d'inclure d'éventuelles EN qui apparaissent entre eux (165). On constate que ces désignateurs sont déterminants pour l'identification de la frontière, puisque leur présence entre l'organisation et le lieu résulte en une scission de l'expression en deux entités (168, à comparer avec 167), opération qui n'a pas lieu en leur absence (169). La majuscule n'est donc pas l'unique critère d'identification des EN.

(164)<ENAMEX TYPE="LOCATION">North Dakota</ENAMEX>
(165)<ENAMEX TYPE="LOCATION">Mississippi River</ENAMEX>
(166)<ENAMEX TYPE="ORGANIZATION">Bridgestone Sports Co.</ENAMEX>
(167)<ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.</ENAMEX>
(168)<ENAMEX TYPE="ORGANIZATION">Hyundai, Inc.</ENAMEX> of <ENAMEX TYPE="LOCATION">Korea</ENAMEX>
(169)<ENAMEX TYPE="ORGANIZATION">McDonald's of Japan</ENAMEX>

Ces exemples nous enseignent également que certaines prépositions ou articles peuvent être inclus dans la séquence de cet identifiant unique (qui contient au moins un mot à majuscule), comme en (167) et en (169). La détermination de l'étendue de ces EN suppose des connaissances sur les propriétés idiosyncratiques des noms propres, c'est-à-dire leur phraséologie : l'exemple (169) semble être une exception car en général deux entités qui se suivent, séparées par une préposition (170), une particule génitive (171), ou même une virgule (172) sont considérées comme distinctes.

(170)<ENAMEX TYPE="ORGANIZATION">University of California</ENAMEX> in <ENAMEX TYPE="LOCATION">Los Angeles</ENAMEX>
(171)<ENAMEX TYPE="LOCATION">California</ENAMEX>'s <ENAMEX TYPE="LOCATION">Silicon Valley</ENAMEX>
(172)<ENAMEX TYPE="LOCATION">Washington</ENAMEX> , <ENAMEX TYPE="LOCATION">D.C.</ENAMEX>

Mais à y regarder de plus près, l'organisation dans l'exemple (170) contient également une préposition, ce qui suppose donc que le nom *University* (qu'on peut considérer comme désignateur) fait partie intrinsèque de la dénomination d'une entité qui s'appelle *University of California*. Il existe en effet une telle entité, elle désigne le pôle universitaire de Californie qui se répartit en 10 universités, dont *UCLA*, ou encore *University of California, Los Angeles*. S'agit-il de ce pôle ou de la branche universitaire du sud de la Californie ? Cet exemple pose le même problème que la firme *McDonald's* et sa branche au Japon en (169) qui sont pourtant associés. On peut donc conclure que les critères de forme l'emportent sur les critères de sens.

La caractérisation sémantique est également affectée par des contraintes discursives qui pèsent sur les phénomènes de reprises anaphoriques. Une EN préalablement introduite comme *Georges W. Bush*, pourra être reprise par le prénom *Georges* ou le nom *Bush*. Si, au sein du texte, *Bush* pourra ne pas être ambigu, il ne constitue pas un identifiant unique de l'entité hors-contexte : on pourra toujours poser la question *Lequel ?*. Le contexte semble donc jouer un rôle dans l'identification du référent lorsque le Npr est « incomplet », soit un identifiant non unique. Le guide recommande pourtant l'annotation de tels Npr. Par exemple, lorsqu'il est en position de modifieur, comme Clinton en (173).

(173)the <ENAMEX TYPE="PERSON">Clinton</ENAMEX> government

On atteint de fortes contradictions lorsque cette même construction est composée de deux EN : un Npr en modifieur et un Npr en tête. Les conventions de MUC-6 indiquaient que lorsque la tête ne doit pas être annotée (ne faisant pas partie des types concernés), le modifieur ne doit pas l'être non plus. Ainsi, dans la séquence *Ford Taurus*, qui désigne des véhicules de marque *Ford* et de série *Taurus*, *Ford* ne doit pas être annoté parce que les produits ou véhicules (*Taurus*) ne font pas partie des EN à annoter. Ce principe sera corrigé dans le guide de MUC-7.

5.2. Des EN plus étendues et plus diverses

Les Entités Nommées se sont progressivement imposées comme une technologie majeure de TAL. Plusieurs campagnes d'évaluation ont pris la relève, comme ACE et CoNLL, et, pour le français, Ester2. Nous n'évoquerons pas CoNLL, car, pour les deux éditions de cette conférence consacrées aux EN (2002 et 2003), la réflexion sur la définition des EN a peu progressé, étant donné que les conventions appliquées étaient celles de MUC-7 [Tjong Kim Sang & De Meulder, 2003]. Dans cette partie, nous aborderons trois problèmes liés à la RCEN : la catégorisation, le discours et les applications.

5.2.1. Ontologies et EN

La tâche d'EN peut être appréhendée comme une tâche de catégorisation sémantique restreinte à certaines expressions linguistiques : les Npr. Nous venons de voir certaines des difficultés rencontrées pour circonscrire les formes que prennent ces catégories dans le cadre de cette tâche et il nous semble que cette tâche gagnerait à être mise en perspective avec le problème de catégorisation sémantique, tel que nous avons pu l'aborder dans les chapitres précédents. À ce compte, nous considérons que les catégories sémantiques désignent des propriétés des référents, que les formes linguistiques permettent de désigner, mais surtout, que ces référents en possèdent une multitude. Si les critères retenus pour les EN et pour les tâches de catégorisation sémantique en général sont d'ordre ontologique (la catégorie Personne est sensiblement similaire à la catégorie Humain), il est théoriquement envisageable d'envisager autant de catégories qu'il existe de dénominations (entendons les syntagmes nominaux). En d'autres termes, il n'y a aucune raison pour que le problème de catégorisation qui s'applique au nom commun ne puisse à présent s'appliquer au nom propre, puisque dans une sémantique référentielle, les Npr désignent (généralement) les référents que des noms communs réunissent selon une propriété [Kleiber, 1996 : 579–584]. Dans ce cadre, la question est de savoir quel type de propriétés on recherche et dans quel but.

La tâche d'EN s'est progressivement décontextualisée de son domaine d'origine, provenant d'une volonté de développer des technologies indépendantes du domaine [Grishman & Sundheim, 1995 : 6], bien que les corpus analysés dans MUC-6 et MUC-7 aient été spécifiques [Grishman, 2010]. La majorité des études ont été effectuées sur des corpus journalistiques qui sont spécifiques du point de vue de leur forme, mais également du point de vue syntaxique et sémantique. L'ontologie devient alors une structure très attractive parce qu'elle est généraliste.

Sekine a par exemple entrepris de réaliser une ontologie d'EN en étendant le nombre de catégories à 150 [Sekine et al., 2002], puis à 200 [Sekine & Nobata, 2004] que nous avons reproduite en figure (5.1).

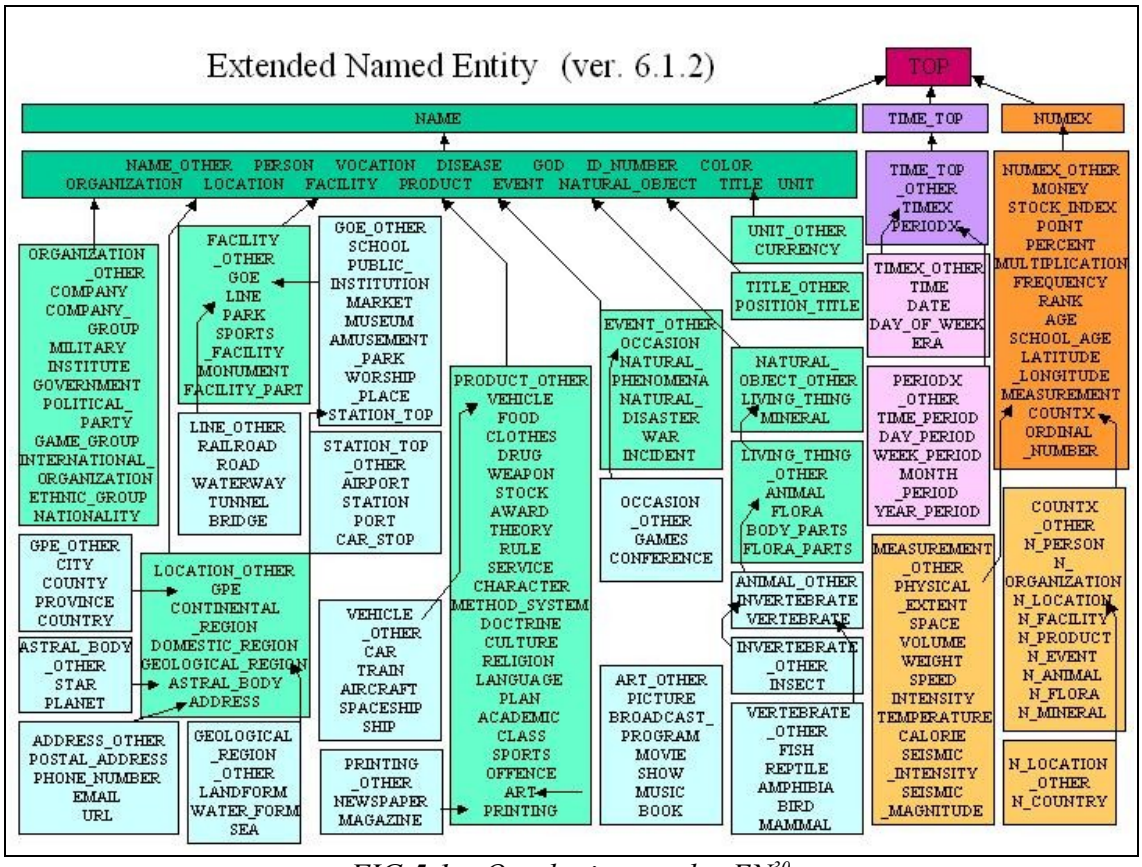


FIG 5.1 – Ontologie pour les EN[30]

Cette classification a été obtenue à partir de trois méthodes : l'annotation manuelle de corpus, la réutilisation de catégories employées dans des systèmes de RCEN et l'extraction automatique de ressources dictionnairiques et encyclopédiques. La définition des EN y est largement étendue à d'autres catégories que les Npr. On y trouve tout d'abord les catégories concernant les unités de mesure (NUMEX) et les expressions temporelles (TIMEX) (qui ne sont pas des ENAMEX, ou *Entity Name Expression*). On y trouve également des formes complexes d'adresse, comme les URL (ADRESS_OTHER). Enfin, il est rare que des catégories comme les sports (PRODUCT_OTHER) ou les titres (TITLE_OTHER) puissent être réalisés par des Npr. Par exemple, les exemples donnés pour la catégorie *Position_Vocation* (comprendre *fonction professionnelle* ; 174) dont le genre est *Title*, ou encore les *Plats*, classés sous *Nourriture* (175) sont rarement des Npr.

(174){suspect, defendant, king, president, CEO, doctor, teacher, Congressman, Foreign Minister, First Lady, manager, coach, captain, player, pitcher}
(175){beer, cola, apple pie, lasagna, General Tso's chicken}

Ni l'unicité du référent ni sa forme ne sont désormais les critères définitoires, la catégorie sémantique l'emporte. On constate le même écart pour le guide d'annotation de la campagne Ester2 [*ESTER 2*]. Cette campagne, appliquée à des corpus de transcription de journaux télévisés en français, proposait sept catégories majeures d'EN et 38 sous-catégories reproduits figure (5.2).

30 http://nlp.cs.nyu.edu/ene/

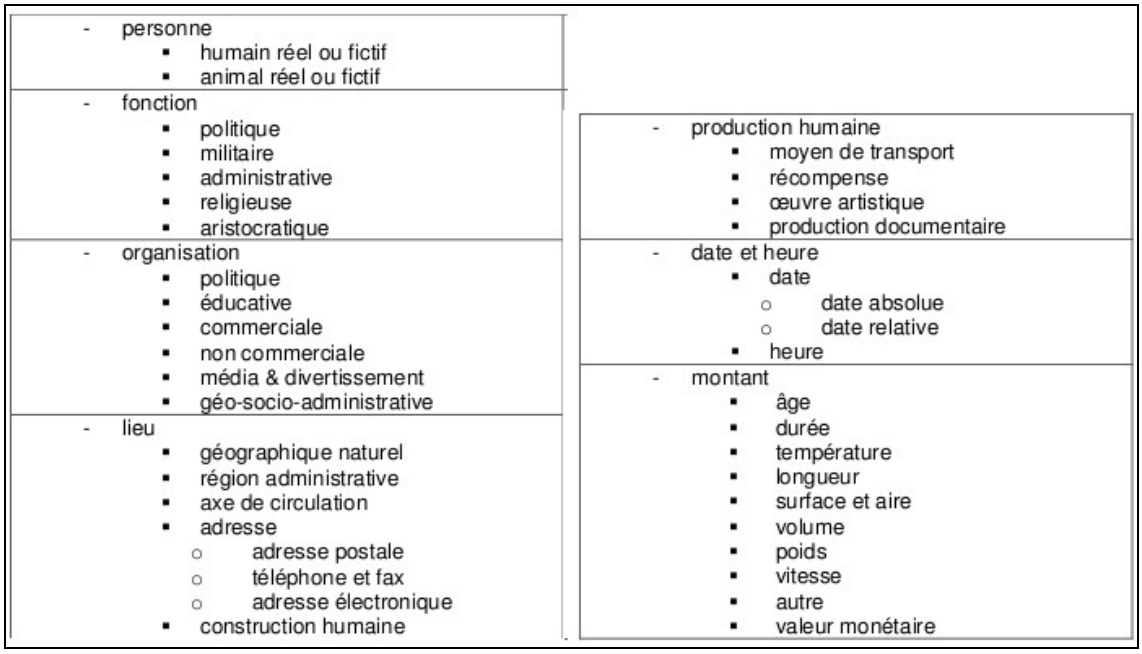


FIG 5.2 – Catégories d'EN de la campagne Ester2 [*ibid.* : 4]

Les conventions d'Ester2 sont hybrides dans le sens où elles n'abandonnent pas tout à fait certains principes de MUC-6 et MUC-7, tout en élargissant la gamme des expressions à annoter. Ce qui justifie la présence de la catégorie *Fonction* par exemple, est la volonté d'inclure certaines expressions référentielles, comme en (176 ; [*ibid.* : 9])

(176)Le [ent=fonc.mil-] commandant [-ent=fonc.mil] a donné son accord pour l'intervention.

Il est évident que dans cet exemple, la forme linguistique qui contribue le plus à la présupposition d'existence de référent est l'article *le*, en emploi spécifique et défini. Le nom commun *commandant* est une des propriétés de ce référent qui le distingue d'autres référents possibles en contexte. Pourtant l'article est exclu de l'annotation. Le Npr n'est plus le désignateur unique et les EN sont définies comme des expressions référentiellement autonomes. Ce critère d'autonomie référentielle est confronté à la variabilité de la présentation d'un référent. La catégorie *Fonction* est par exemple annotée même si le nom n'est pas précédé d'un article spécifique défini, comme lorsqu'il est exclu dans une apposition (177 ; [*ibid.*]).

(177)[ent=pers.hum-] [ent=pers.hum-] Pierre Lellouche [-ent=pers.hum], [ent=fonc.pol-] député UMP de [ent=loc.admi-] Paris [-ent=loc.admi] [-ent=fonc.pol] [-ent=pers.hum], estime dans un entretien ...

À côté de tels exemples, le guide d'Ester2 conserve la convention de MUC-7, consistant à ne pas annoter les « civilités ». Il en fournit une liste :

> « Ne pas annoter les civilités, autres que les fonctions, qui complètent les entités nommées.
> Il s'agit de Monsieur, Madame, Mademoiselle, Maître, Docteur, Professeur » [*ibid.* : 8]

Ce choix laisse dubitatif : *Docteur* et *Professeur* ne sont-ils pas sémantiquement des noms de fonction, comme peut l'être *comtesse* (comparer 178 et 179) ? Voudrait-on véritablement qu'un système automatique puisse discriminer de tels usages ?

(178)J'ai été opéré par le Professeur [ent=pers.hum-] Causset [-ent=pers.hum]. [*ibid.*]
(179)La [ent=pers.hum-] [ent=fonc.ari-] comtesse [-ent=fonc.ari] [ent=pers.hum-] de Ségur [- ent=pers.hum] [-ent=pers.hum] est une femme de lettres française [*ibid.* : 10]

Il est également possible de trouver *professeur* ou *docteur* en apposition, auquel cas on peut supposer qu'il serait annoté comme *Fonction*. La distinction entre la catégorie Civilité et Fonction semble donc résider dans la position (les civilités précèdent les Npr de Personne) et, peut-être, la majuscule (*Professeur* en 178).

Une des solutions serait d'envisager des guides entièrement établis sur des critères sémantiques. La campagne ACE [Doddington et al., 2004] est à cet égard exemplaire puisqu'elle proposait d'annoter la totalité des expressions linguistiques, pronoms personnels, pronoms interrogatifs et relatifs, Npr, expressions référentielles. Néanmoins la complexité des phénomènes en jeu a des conséquences inéluctables sur la capacité des systèmes à reconnaître et à classifier correctement toutes ces expressions, notamment parce qu'ils doivent prendre en compte la totalité du document (chaînes anaphoriques par exemple).

5.2.2. Le discours : l'EN en contexte

Si la tâche d'EN a pu paraître simple dès l'origine, comparée à des tâches comme l'extraction de relations au sein d'un cadre, elle s'avère bien plus complexe lorsqu'on souhaite modéliser la sémantique des EN en contexte.

Le premier problème est celui de la polycatégorialité, ou homonymie. Les tâches d'annotation menées lors des différentes campagnes d'évaluation ont fait émerger le fait qu'un même Npr pouvait être classé dans plusieurs catégories. Le cas du Npr *Charles de Gaulle* est exemplaire comme l'illustrent les exemples (180), (181) et (182) [ESTER2, 2007 : 5].

(180)[ent=pers.hum-] Charles de Gaulle [-ent=pers.hum] est le fondateur de la Cinquième République.
(181)Le [ent=prod.vehicule-] Charles de Gaulle [-ent=prod.vehicule] a connu son premier jour de mer le 27 janvier 1999.
(182)Mon avion arrive demain à [ent=loc.fac-] Charles de Gaulle [-ent=loc.fac].

En (180), ce Npr désigne une personne, en (181) un porte-avion, en (182) un aéroport. Il est également fréquent que des organisations soient nommées d'après le nom d'une personne. Que l'on songe aux noms d'entreprise de mode comme *Calvin Klein* par exemple, à certaines sociétés automobiles comme *Renault*, *Mercedes*. Les noms de pays figurent aussi dans les Npr d'organisation ou de journaux, comme *France Loisirs*, *Air France*, *France 5*, *France Soir*, etc. Les Npr de personnes peuvent provenir de noms de lieux, comme *Aigremont*, *Roubaix*, etc. Enfin, les noms de lieux peuvent provenir de noms de personnes (noms de rues par exemple). En revanche, les noms d'organisation donnent rarement lieu à des noms de personne, mais plutôt à des noms de produits comme la boisson *Coca*, ou les voitures *Ford*. L'identité graphique n'est donc pas garante d'une stabilité référentielle. La catégorisation des EN est donc une tâche complexe, les connaissances que l'on peut avoir acquises par ailleurs ne suffisent pas, et une analyse du contexte et de l'usage de ces expressions s'imposent.

5.2.Des EN plus étendues et plus diverses

Au phénomène que nous avons nommé la polycatégorialité, s'ajoute un second d'ordre discursif, la métonymie, que nous avons eu l'occasion d'aborder dans le cadre des alternances sémantiques (cf. *infra* chapitre 4) et de la métonymie intégrée (cf. *infra* 1.2) : les Npr, comme les noms communs peuvent référer à une autre entité que celle qu'ils désignent généralement (*Maison Blanche*, *Matignon*, *Place Beauvau*, etc.). La différence fondamentale avec la polycatégorialité est que la catégorie de l'entité employée métonymiquement est reconnue et détermine son interprétation métonymique : il y a un rapport sémantique entre les deux catégories. Par exemple, les noms de lieux (notamment de pays et de villes) peuvent être employés dans des contextes autres que la localisation. En (183), le véritable sujet du verbe *gagner* serait plutôt un collectif de personnes qu'un lieu (une équipe de sport qui représente ce pays par exemple).

(183)La [ent=org.div-] France [-ent=org.div] a gagné en finale. [*ibid.*]

Outre le fait que *France* a ici un rôle syntaxique discriminant qui puisse justifier un tel transfert, il semble que les noms de pays comme *France* [Cruse, 1996] soient des entités complexes à plusieurs facettes (territoire, gouvernement, habitants, etc.), autrement dit, qu'ils « intègrent » plusieurs entités. La métonymie peut donc se comprendre comme l'usage d'une facette plutôt que d'une autre, plus saillante. La catégorie *Organisation géo-politique* est parfois proposée pour les traiter séparément ([Sekine et al., 2002] ; voir les catégories *GPE* et *GPE_other* dans la figure 5.1, p143). Dans l'exemple (183) la métonymie de *France* est directement résolue : c'est une organisation.

Pour ce qui concerne la campagne MUC (et donc les tâches de CoNLL), le problème de métonymie était explicitement écarté (sauf quelques particuliers dits idiosyncratiques comme *Maison Blanche*). En revanche l'ambition d'Ester2 était une annotation contextuelle : les annotateurs avaient la difficile tâche de résoudre les métonymies en tenant compte « de l'intention de l'auteur » [ESTER2, 2007 : 5]. Comme on pourra s'en rendre compte à travers les exemples (184) et (185), tirés du guide d'annotation, la tâche peut mener à des contradictions (voir aussi [Poibeau, 2006 : 1965]).

(184)Je me suis fait opér[er] à l'[ent=org.non-profit-] hôpital Necker [-ent=org.non-profit]. [ESTER2, 2007 : 11]
(185)Je me suis fait opér[er] à l'[ent=loc.fac-] hôpital Broussais [-ent=loc.fac]. [*ibid.* : 6]

La métonymie des EN a fait l'objet d'une tâche spécifique dans le cadre de la campagne Semeval-7 [Markert & Nissim, 2007], ce qui a permis de donner un premier éclairage qualitatif et quantitatif à ce problème. L'objectif était de distinguer les différentes interprétations possibles des noms de lieux et des noms d'organisation. Les systèmes pouvaient être évalués sur trois niveaux de granularité pour chacune de ces catégories (tableaux 5.2 et 5.3).

Niveau	Catégorie	Fréquence	Proportion
Coarse	Litteral	721	0,79
	Non-Litteral	187	0,21
Medium	Litteral	721	0,79
	Metonymic	167	0,18
	Mixed	20	0,02
Fine	Litteral	721	0,79
	Place-for-People	141	0,16
	Mixed	20	0,02
	Othermet	11	0,01
	Place-for-Event	10	0,01
	Object-for-Name	4	0,0004
	Place-for-Product	1	0,0001
	Object-for-Representation	0	0

Tableau 5.2 – Métonymies de noms de Lieux de Semeval-7 [*ibid.* : 39]

Niveau	Catégorie	Fréquence	Proportion
Coarse	Litteral	520	0,62
	Non-Litteral	322	0,38
Medium	Litteral	520	0,62
	Metonymic	262	0,31
	Mixed	60	0,07
Fine	Litteral	520	0,62
	Organisation-for-Members	161	0,19
	Organisation-for-Product	67	0,08
	Mixed	60	0,07
	Organisation-for-facility	16	0,02
	Othermet	8	0,01
	Object-for-Name	6	0,007
	Organisation-for-Index	3	0,003
	Organisation-for-Event	1	0,001
	Object-for-representation	0	0

Tableau 5.3 – Métonymies de noms d'Organisations de Semeval-7 [*ibid.*]

Les trois niveaux sont « coarse », « medium » et « fine » et désignent respectivement :

- Au premier niveau, on cherche uniquement à savoir si l'expression est littérale ou pas.
- Au second niveau, la catégorie mixte est introduite pour les cas où une séquence revêt les deux interprétations à la fois.
- Au troisième niveau, 14 catégories supplémentaires sont introduites (8 pour les organisations et 6 pour les lieux) et correspondent aux types de transfert métonymiques documentés dans [Markert & Nissim, 2003]

En termes de fréquence (il s'agit des caractéristiques du corpus de test, sur lequel sont évalués les systèmes), on remarque que, pour les lieux, la catégorie littérale (0,79) est sur-représentée, suivie par la catégorie *Place-for-people* (0,16). Cette dernière catégorie regroupe les cas où les lieux

désignent par métonymie des personnes et ceux où ils désignent des organisations représentatives de ce lieu, comme en (183). Pour les organisations, il s'agit également de la catégorie littérale (0,62), puis « Organisation-for-members » (0,19) et « Organisation-for-Product » (0,8), soit 90% des données de test. À titre illustratif, nous donnons les résultats pour le système XRCE-M [Brun et al., 2007] dans les tableaux (5.4) et (5.5) pour les lieux et les organisations respectivement.

	Nb occ.	Prec.	Recall	F-score
Literal	721	0.867	0.960	0.911
Place-for-people	141	0.651	0.490	0.559
Place-for-event	10	0.5	0.1	0.166
Place-for-product	1	_	0	0
Object-for-name	4	1	0.5	0.666
Object-for-representation	0	_	_	_
Othermet	11	_	0	0
mixed	20	_	0	0

Tableau 5.4 – Résultats du système Xrce-M pour les lieux [*ibid.* : 491]

	Nb occ.	Prec.	Recall	F-score
Literal	520	0.730	0.906	0.808
Organization-for-members	161	0.622	0.522	0.568
Organization-for-event	1	_	0	0
Organization-for-product	67	0.550	0.418	0.475
Organization-for-facility	16	0.5	0.125	0.2
Organization-for-index	3	_	0	0
Object-for-name	6	1	0.666	0.8
Othermet	8	_	0	0
Mixed	60	_	0	0

Tableau 5.5 – Résultats du système Xrce-M pour les organisations [*ibid.*]

On constate que la fréquence est souvent corrélée avec la performance du système. L'identification des lectures métonymiques est problématique même si elles sont suffisamment représentées par rapport à la catégorie *Literal* : la détection de la métonymie *Lieu→Personne* atteint 0,55 de F-mesure, celle de *Organisation→Membres*, 0,56, et celle de *Organisation→Produit*, 0,47. Le problème pour interpréter ces résultats est que l'on ne sait s'il faut attribuer la baisse de performances aux systèmes ou aux conventions d'annotation.

Toutes ces difficultés de catégorisation sémantique nous ramènent à la question des critères qui prévalent à la définition d'une catégorie. Le contexte d'application y joue un rôle déterminant.

5.2.3. Contexte d'application et EN

L'application a une influence majeure sur la conception des systèmes de RCEN, elle définit le contexte qu'il doit modéliser : le domaine de connaissances et l'objectif qu'on lui assigne. Comme nous l'avons vu, ces systèmes ont été conçus au départ pour répondre à un objectif de généralité et d'autonomie, mais c'était sans compter sur le domaine. Dans cette perspective, nous considèrerons deux applications majeures de ces systèmes que nous tenterons de rapprocher : l'EI et la Recherche d'Information Précise.

Un des objectifs de MUC était d'identifier les noms propres apparaissant dans des scénarios précis. Le domaine de connaissance était donc délimité par ces scénarios. Les corpus devaient aussi être adaptés à la tâche (les scénarios doivent y figurer, au moins en partie). À ce propos, Sundheim nous rappelle que les corpus employés créaient un « biais » du point de vue du style et du sujet :

> « It represents just one style of writing (journalistic) and has a basic [bias] toward financial news and a specific bias toward the topic of the Scenario Template task. » [Sundheim, 1995 : 16]

Autrement dit, les systèmes développés sont valables pour tel style de corpus (presse financière) et tel sujet mais on n'a aucune idée de leurs performances dans des corpus de style ou de sujet différent. En effet, dans cette campagne, les textes étaient collectés grâce à un système de RI et à partir de requêtes de mots-clés. La présence de certains mots dans le texte, comme *assassination*, *kidnapping* ([*ibid.* : 14] ; [Sundheim, 1991 : 6]) permettait dans une certaine mesure de contrôler le sujet. Les corpus étaient pour la majorité des extraits du Wall Street Journal.

Une des questions légitimes est donc d'évaluer la portabilité des systèmes de RCEN dans des corpus qui relèvent de domaines généraux et dont le style est sensiblement similaire. C. Mota et R. Grishman [Mota & Grishman, 2008] rappellent d'ailleurs que dans les protocoles d'évaluation de la campagne MUC, les corpus sur lesquels les systèmes ont été développés sont de même nature que les corpus sur lesquels ils sont évalués, ce qui confirme d'une certaine manière, leur spécificités. Leur objectif est d'évaluer les performances de leur système en fonction du critère de variation temporelle du corpus de test. Ils montrent ainsi que la quantité d'EN communes et les performances des systèmes chutent en fonction de la distance temporelle entre le corpus de développement et le corpus de test (moins de dix ans). T. Poibeau et L. Kosseim [Kosseim & Poibeau, 2001], ont pour leur part documenté les conséquences d'un changement de type de texte sur les performances de leurs systèmes. L'évaluation a porté sur un corpus de transcription téléphoniques de qualité et un corpus d'échange de courriels. Comparé à l'évaluation MUC sur des corpus journalistiques, la F-mesure des quatre systèmes étudiés chute de manière importante (plus de 40%). Un cas spécifique est celui des retranscriptions de documents audios ou de textes reconnus par OCR pour lesquels la qualité du texte a un impact indiscutable sur les performances des systèmes ([Galliano et al., 2009] ; [Galibert et al., 2010]). La robustesse des systèmes de RCEN peut ainsi être mise à l'épreuve par de nombreux facteurs.

Mais on peut considérer le contexte d'application qu'est l'EI financière comme un avantage : il limite la complexité du vocabulaire à traiter ainsi que les types d'EN à identifier. Par exemple, on peut faire l'hypothèse que des organisations comme les clubs sportifs, les hôpitaux, les médias, les entreprises de mode n'apparaîtront jamais dans un texte relevant d'un scénario comme l'achat d'aéronefs. Le scénario permet également de définir les rôles et les relations auxquels un certain type d'EN est associée : en simplifiant, le scénario de succession de poste, par exemple, met en relation une EN de type Personne avec une EN de type Organisation et les rôles correspondent par exemple à la personne destituée ou au remplaçant, à l'ancienne ou à la nouvelle organisation d'affiliation. On peut donc croire qu'une meilleure définition des relations auxquelles est associée une EN peut garantir leur catégorisation.

La tâche de détection de relations entre EN a été isolée et définie officiellement dans l'édition 2002 de la campagne ACE. Elle renverse la perspective en se focalisant, non plus sur les scénarios, mais sur les EN, tout en conservant un aspect du scénario : la relation. Les cinq types de relations à identifier associaient les trois EN majeures [ACE2002, 2002 : 3] :

- Le rôle (Role) : désigne l'affiliation d'une personne à une organisation.
- La partie (Part) : désigne les relations de dépendances entre organisations
- La localisation exacte (At) : désigne le lien entre une personne ou une organisation et un lieu
- La localisation approximative (Near) : indique que le lien de localisation est proche
- Les relations sociales (Social) : désigne les liens entre personnes du point de vue personnel et professionnel

Chacune de ces relations peut être distinguée en sous-types, comme les relations sociales en relations familiales (filiation, mariage, etc.) et relations professionnelles (patron, collègue, etc.). Dans ce cas, la majorité des relations sont relatives au scénario de changement de poste.

Les systèmes de RCEN sont également utilisés dans les systèmes de Question-Réponse (SQR) à domaine ouvert. Ces systèmes doivent pouvoir fournir une réponse à une question posée en langue naturelle. Cela suppose principalement une analyse de la question, une recherche de documents pertinents (RI) et une extraction de la réponse. Les systèmes de RCEN y sont employés pour l'extraction de la réponse dans les documents (les EN peuvent être identifiées à la volée ou préalablement), mais sont liés à l'analyse de la question : l'analyse de certaines questions (*Où est Charlie ?*) suppose d'identifier le type de réponse attendu (EAT ou « Expected Answer Type »), qui lorsque c'est possible, correspond à une catégorie sémantique (potentiellement une EN ; [Molla et al., 2006]). Si le type est correctement identifié (l'une des causes d'erreurs majeures ; [Moldovan et al., 2003 : 143]), le SQR pourra sélectionner les passages dans les documents qui contiennent cette catégorie sémantique (ainsi que d'autres informations pertinentes de la question). On peut supposer que le EAT d'une question débutant par *où* aura de fortes chances d'être un Lieu par exemple. Mais, comme l'ont observé D. Radev et ses collègues [Radev et al., 2002], les pronoms interrogatifs ne sont pas garants d'un EAT correct. Ces derniers ont analysé près de 500 questions pour documenter la variabilité des types (sémantiques) attendus en fonction des pronoms interrogatifs, illustrés dans le tableau (5.6).

Wh-word	Types
who (102)	PERSON(77) DESCRIPTION(19) ORG(6)
where(60)	PLACE(54) NOMINAL(4) ORG(2)
when(40)	DATE(40)
why(1)	REASON(1)
what /which(233)	NOMINAL(78) PLACE(27) DEFINITION(26) PERSON(18) ORG(16) NUMBER(14) ABBREVIATION(13) DATE(11) RATE(4) KNOWNFOR(8) MONEY(3) PURPOSE(2) REASON(1) TRANSL(1) LENGTH(1) DESCOTHER(10)
how(48)	NUMBER(33) LENGTH(6) RATE(2) MONEY(2) DURATION(3) REASON(1) DESCOTHER(1)

Tableau 5.6 – Type de réponse attendue en fonction des pronoms interrogatifs [*ibid.*]

Si l'on s'intéresse à présent à l'extraction de la réponse, nous avons vu que les EN étaient sujettes à des phénomènes de variation de sens en contexte : la qualité d'un système de RCEN peut influencer (positivement ou négativement) les performances d'un SQR. Par exemple, si l'EAT est de type Organisation (*Qui a remporté la coupe du monde de football en 1998 ?*), mais que le système de RCEN a mal désambiguïsé une entité de Lieu en emploi métonymique (*France)*, le SQR pourra être induit en erreur en ne sélectionnant pas les passages appropriés. À notre connaissance aucun SQR ne traite spécifiquement de la métonymie ; pour gérer ce problème en général, les types à rechercher sont multipliés, comme le proposent Mollá et ses collègues [Molla et al., 2006 : 53]. Ils rappellent en effet que le module d'EN est souvent développé indépendamment du SQR et des desiderata de la tâche.

L'adaptation d'un système de RCEN aux besoins d'un SQR constitue donc une piste de recherche pour améliorer les performances de ce dernier. En étudiant les questions, on peut déterminer le EAT et envisager une annotation d'EN dans le corpus en fonction. Néanmoins, les questions auxquelles sont soumis les SQR peuvent être extrêmement diverses [Voorhees, 2000 : 1] et la base documentaire dont dispose le SQR aussi.

Les campagnes d'évaluation menées sur les SQR (TREC[31], QA@CLEF[32], EQUER[33], QAST[34]) ont permis de définir des catégories majeures de question. Pour illustrer cette variété, nous nous appuierons uniquement sur la campagne TREC et nous distinguerons quatre temps (pour une présentation détaillée, voir [Galibert, 2009 : 124]).

- Les questions proposées au départ (TREC-8) étaient de type « factoïde » (encore appelé « trivial ») : les distinctions étaient explicitement établies en fonction du type d'EN recherché dans la question (personne, lieu, date, montant, etc.). Les questions ciblaient les informations typiquement recherchées sur ces entités (naissance ou profession, âge à telle date pour une personne par exemple). Ce sont généralement les questions les plus fréquentes.
- La campagne TREC-10 a introduit la tâche de listes (*nommer 4 ouvrages de Stevenson*) ainsi que la tâche contextuelle, dans laquelle plusieurs questions sont regroupées autour d'une même entité (tâche qui sera abandonnée).
- La campagne TREC-13 a adopté un nouveau format, en reprenant la tâche contextuelle sous le nom de « question series », regroupées autour d'une cible commune (« target ») comportant quelques questions factoïdes, quelques listes et une question OTHER (information supplémentaire). La nouveauté de cette édition était également d'introduire les événements (par exemple, le « Teapot Dome scandal ») parmi les types de catégories recherchées.
- Le dernier type de question proposé par les campagnes d'évaluation TREC est nommé les « questions de relations complexes » (Introduites dès TREC-14). Les relations envisagées étaient énumérées ainsi : relations financières, mouvements de biens, liens familiaux, voies de communication, liens organisationnels, co-location, intérêts communs, et relations temporelles. Les questions sont, d'un certain point de vue, moins précises, car plusieurs réponses peuvent être apportées (à la différence de questions factoïdes). Voici quelques exemples :

31 http://trec.nist.gov/
32 http://clef-qa.fbk.eu/index.html
33 http://www.technolangue.net/article195.html
34 http://www.lsi.upc.edu/~qast/2007/

(186)What [familial ties] exist between [dinosaurs] and [birds]?
(187)What [financial relationships] exist between [the Israeli government] and [the Palestinian National Authority (PNA)]?
(188)What effect does [second-hand smoke] have on [non-smokers]?
(189)What is the position of [the Saudi Government] with respect to [Osama bin Laden]?
(190)Is there evidence to support the involvement of [the North Korean Government] in [currency counterfeiting]?

Les questions factoïdes semblent donc être celles qui sont directement pertinentes vis-à-vis des systèmes des RCEN. Dans ce cas, la phase d'analyse de la question peut être conçue comme l'identification d'un scénario de recherche d'information qui se matérialise en corpus par l'expression d'une relation à laquelle est associée une EN. Autrement dit, la tâche d'analyse de question est fortement similaire à la définition de scénarios en EI [Srihari & Li, 2000].

La question du contexte d'application d'un système de RCEN soulève bon nombre de questions et nourrit d'intéressantes pistes de recherche. Il existe différents types de systèmes de RCEN et la problématique de l'adaptation dépend de la technologie adoptée.

5.3. Les systèmes de RCEN

La forme la plus simple d'un système de RCEN consiste à projeter un lexique d'EN d'une catégorie donnée sur un corpus : si une séquence de caractères alphanumériques est reconnue, elle est annotée par cette catégorie. Un tel système réduit une catégorie sémantique à sa forme graphique, ce qui n'est pas satisfaisant, au vu de ce que nous avons pu observer en 5.2. Bien qu'ils posent des inconvénients en terme de couverture (un lexique doit être mis à jour) et d'absence d'analyse contextuelle ([McDonald, 1996] ; [Mikheev et al., 1999]), les lexiques restent néanmoins une source de connaissance critique avec les corpus annotés. En s'appuyant sur ces ressources, les systèmes qui ont été développés pour la RCEN sont principalement de deux types : systèmes symboliques à base de règles et classifieurs statistiques.

5.3.1. Les systèmes symboliques

Les systèmes symboliques modélisent les connaissances requises pour détecter et catégoriser une EN. Ils peuvent être distingués en fonction du degré de complexité de la grammaire qu'ils implémentent et des types de ressources employées. Parmi les lexiques, on peut distinguer les lexiques sémantiques suivants :

- lexiques de Npr (« gazeteer »)
- lexiques de mots déclencheurs (« trigger-words »)
- lexiques encyclopédiques (Wordnet)
- bases de données (DBpédia[35])

La majorité des systèmes symboliques utilisent des grammaires régulières et définissent des patrons lexico-syntaxiques permettant de reconnaître des motifs dans lesquels apparaissent les EN. Ces patrons sont employés pour corriger la catégorie d'une EN d'un lexique à partir d'indices contextuels et pour reconnaître des EN inconnues des lexiques. La figure (5.3) illustre un automate simple de détection d'une Personne, comme *Dr. Watson*.

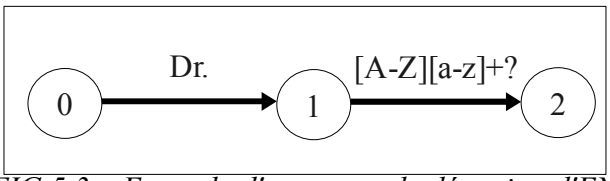


FIG 5.3 – Exemple d'automate de détection d'EN

Dans cet automate, on peut substituer *Dr.* par une liste de mots déclencheurs jouant un rôle identique lorsqu'ils sont suivis d'un mot à majuscule. R. Magnini et ses collègues [Magnini et al., 2002] proposent par exemple d'extraire de telles listes à partir des catégories de Wordnet[36] (lexique encyclopédique).

Les grammaires modélisent le contexte de manière linéaire (progression gauche-droite ou droite-gauche) et l'information employée est généralement locale (au sein du syntagme). Elles sont, comme les lexiques, sujettes à des limites de complétude car elles supposent une analyse fine et

35 http://dbpedia.org
36 http://wordnet.princeton.edu/

exhaustive du contexte dans lequel apparaissent les EN ciblées. La complexité de ces grammaires peut être illustrée par la figure (5.4) qui montre un transducteur permettant de détecter des noms d'organisations de type *Établissement scolaire* (extrait de Friburger, 2002 : 88).

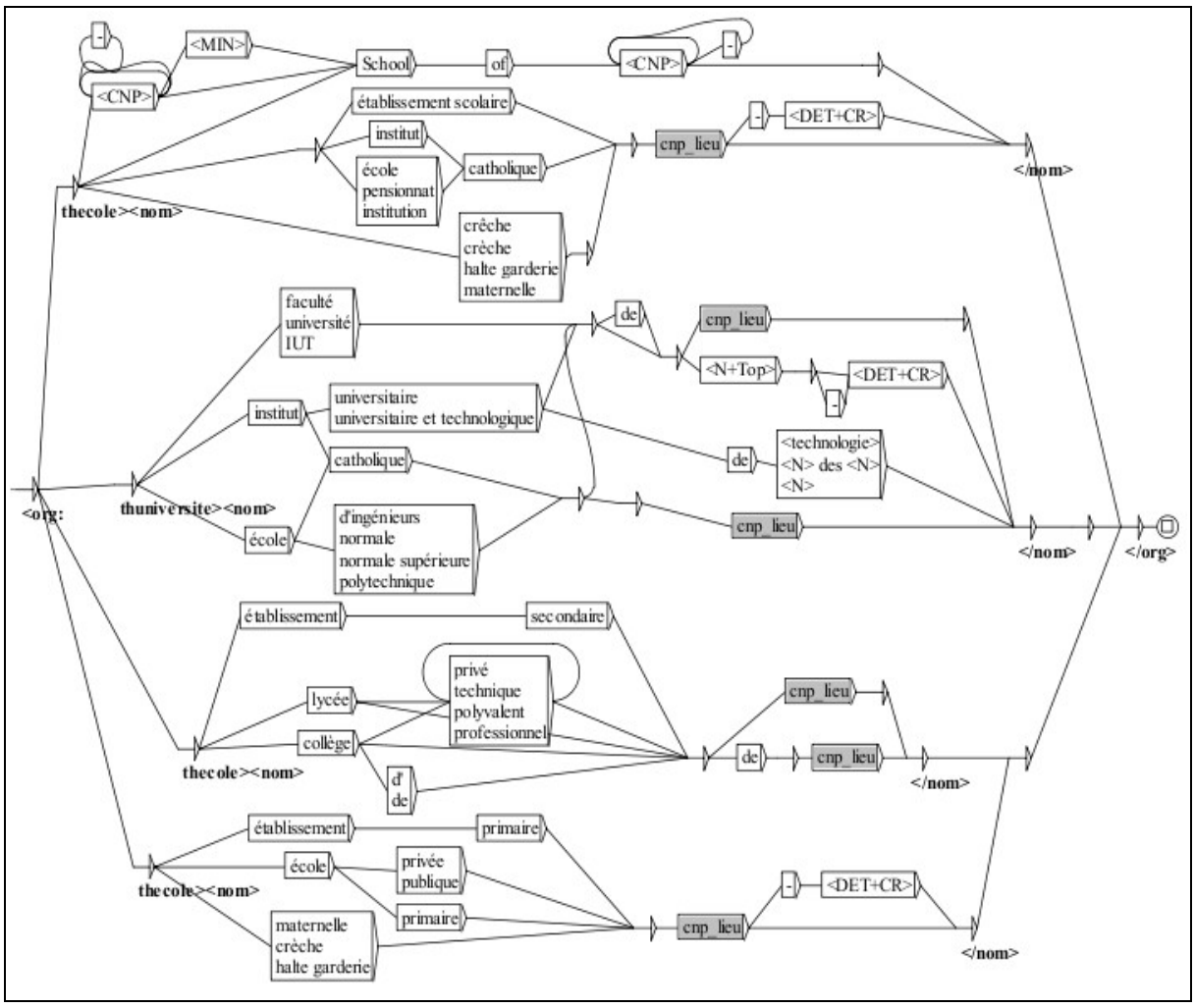


FIG 5.4 – Exemple de transducteur issu de CasSys pour la reconnaissance d'établissement scolaire

Ce transducteur est issu du système CasSys [Friburger, 2002], basé sur la plateforme Intex [Silberztein, 1993]. Il se lit de gauche à droite et insère au fur et à mesure les balises des entités reconnues (en gras). Par exemple, l'identification de la séquence *la crèche de Lorient* commencera par ***<org:*** puis ***thecole><nom*** ; le mot *crèche* (le déclencheur de la règle) permettra de parvenir à un autre transducteur (cnp_lieu en gris), qui confirmera *Lorient* comme lieu et la boucle s'achèvera par ***</nom>*** et ***</org>*** donnant un résultat possible, illustré en (189).

(191)*la* ***<org:thecole> <nom>*** *crèche de* ***<lieu:top><nom>****Lorient****</nom></lieu></nom></org>***

Les transducteurs peuvent s'appuyer sur les formes exactes ou sur des propriétés définies dans des lexiques (commes les catégories morpho-syntaxiques *DET+CR* : déterminant ou chiffre romain ; *N+Top* : toponyme). Le graphe montre que la détection des établissements scolaires repose sur des déclencheurs comme *école*, *institut* ou *établissement*. Les états font correspondre des séquences

textuelles occupant la même place dans la dénomination bien que leur nature syntaxique puisse différer (comparer *d'ingénieurs* et *normale supérieure*). Pour être efficaces, ces grammaires locales doivent minimiser le nombre d'états et de transitions et ainsi éviter la dispersion et la multiplication des lexiques (modularité). Ce travail suppose une expertise linguistique pour une tâche qui peut s'avérer lourde en temps de traitement, notamment dans le cas d'une adaptation à de nouveaux textes, styles ou domaines.

Étant locale, une telle analyse ne peut à elle seule correctement assigner les catégories sémantiques issues de phénomènes métonymiques : leur gestion suppose souvent des traitements de haut-niveau exploitant un contexte plus large. C'est d'ailleurs l'une des principales sources d'erreur du système CasEN, distribution récente adaptée de CasSys, évalué dans le cadre de la campagne Ester2 [Nouvel et al., 2010]. En effet, la détection de métonymie peut s'appuyer par exemple sur la nature du verbe dont l'EN est le sujet (verbe de cognition ou de parole), qui suppose une analyse syntaxique. C'est le traitement proposé par C. Brun et ses collègues [Brun et al., 2007] dans le cadre de leur travail d'adaptation d'un système de RCEN sur l'anglais (les performances sont présentées en 5.2.2). Les règles qu'ils proposent s'appuient sur les relations de dépendances syntaxiques fournies par l'analyseur XIP [Aït-Mokhtar et al., 2002], sur des listes de verbes ainsi que sur les catégories du système de RCEN initial. Par exemple, certaines métonymies *Lieu-pour-Personne*, comme en (192), peuvent être détectées lorsqu'un nom de pays est sujet d'un verbe d'action économique (comme *payer*).

(192)La France va devoir payer des amendes considérables (Web)

L'adaptation des systèmes symboliques est une tâche complexe qui dépend de la formalisation adoptée par le système de base [Chiticariu et al., 2010].

5.3.2. Systèmes supervisés

La tâche d'EN a très tôt intéressé la communauté d'apprentissage automatique (Machine Learning) dépassant ainsi le cadre originel du TAL. La conférence CoNLL, dédiée à l'application de techniques d'apprentissage automatique pour le traitement de la langue, a proposé deux campagnes d'EN pendant lesquelles les techniques d'apprentissage supervisé (AS) ont pu être confrontées. L'apprentissage est dit supervisé parce que les classes sont données par avance (leur nature et leur nombre) sous forme de corpus annoté (le corpus d'apprentissage). Plus spécifiquement, on considère une approche supervisée lorsque le système dispose en entrée de données de référence qu'il doit répliquer. À l'inverse, l'apprentissage non-supervisé (ANS) fait émerger des classes ou clusters de données ayant des structures similaires, comme on peut le faire avec la classification ascendante hiérarchique (cf. *infra* chapitre 3).

La tâche d'EN est alors conçue comme un problème de classification statistique dans lequel des ensembles de traits (« feature set ») extraits à partir d'exemples (l'entrée) prédisent une classe d'EN (la sortie). Ces systèmes assignent un score de vraisemblance ou de probabilité à l'une des classes pour chaque trait. Le modèle statistique obtenu est alors utilisé pour répliquer l'annotation sur des corpus similaires et notamment sur le corpus de test par rapport auquel il est évalué.

Les systèmes supervisés peuvent principalement se distinguer en fonction de la construction des ensembles de traits et de l'algorithme de classification.

- Les ensembles de traits constituent l'essentiel de l'information extraite des exemples d'apprentissage. Les traits correspondent à des catégories de mot, comme sa catégorie morpho-syntaxique. Plus généralement, les traits correspondent à des propriétés de mots,

comme le fait qu'ils possèdent une majuscule initiale, qu'ils soient en début de phrase ou encore qu'ils soient un nombre à quatre chiffres. La figure (5.5) résume par exemple les traits utilisés par Bikel et ses collègues [Bikel et al., 1999] pour créer un modèle statistique basé sur un modèle de Markov caché (HMM).

Word Feature	Example Text	Intuition
twoDigitNum	90	Two-digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11/9/89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount, percentage
otherNum	456789	Other number
allCaps	BBN	Organization
capPeriod	M.	Person name initial
firstWord	*first word of sentence*	No useful capitalization information
initCap	Sally	Capitalized word
lowerCase	can	Uncapitalized word
other	,	Punctuation marks, all other words

FIG 5.5 – Traits employés pour le modèle HMM de [*ibid.*]

Comme le montre la figure, les traits correspondent à des intuitions jugées pertinentes pour catégoriser une EN : l'intérêt de créer le trait *allCaps* est qu'il peut être pertinent pour discriminer les organisations. Ces traits sont binaires (0 ou 1) et caractérisent autant le mot évalué que son contexte (les mots qui l'environnent). Lorsque ce n'est pas le cas (comme pour l'annotation morpho-syntaxique ; ce qui n'est pas le cas dans la figure 5.5), les données peuvent être décomposées (*est_un_nom*, *est_un_adjectif*, etc.). D'autres types de traits sont employés, comme les collocations ou n-grammes apparaissant dans une fenêtre de taille arbitraire, la présence d'un mot dans un lexique, ou des combinaisons de ces traits (catégorie + position), voire des sorties de système de RCEN. Les structures linguistiques sont relativement peu employées (mais voir [Benajiba et al., 2010] pour un usage de la syntaxe).

- Parmi les algorithmes employés, on distingue les approches génératives et discriminantes. Les méthodes discriminantes cherchent à prédire à quelle classe y appartient un échantillon x (mot) en évaluant directement le pouvoir discriminant de ses traits (distribution de la probabilité a posteriori p(y|x)) : les machines à vecteur support (SVM ; [Isozaki & Kazawa, 2002]), la discrimination à entropie maximale [Borthwick, 1999] et les arbres de décision [Sekine et al., 1998] sont les plus fréquemment employés pour la RCEN ; Les CRF (champs aléatoires conditionnels ; [Lafferty et al., 2001]), sont actuellement les modèles qui obtiennent les meilleurs performances. Les méthodes génératives établissent la probabilité qu'un échantillon x appartienne à une classe y en modélisant les dépendances entre toutes les variables du modèle (p(x,y)), pour en déduire le pouvoir discriminant d'un trait. La méthode la plus couramment employée pour la RCEN est le modèle de Markov caché (HMM ; [Bikel et al., 1999]). Pour résumer, la probabilité a posteriori qu'un mot appartienne à une classe d'EN donnée est obtenue par le produit de toutes les probabilités entre les variables du modèle (probabilités de transition ou événements) qui apparaissent dans un espace de texte donné (bigrammes avec le mot à gauche ou à droite pour [*ibid.*]) : elle est induite.

Les systèmes supervisés réduisent considérablement l'effort de conception de systèmes de RCEN. Leur intérêt est que le cœur du système (mis à part la définition des traits et le choix de la taille du contexte) ne change pas, ce qui facilite leur portabilité à d'autres langues ou domaines. De plus, ces systèmes nécessitent des corpus d'entraînement suffisamment larges et de qualité pour être performants. Le problème qui se pose n'est plus le temps de création de règles comme pour les systèmes symboliques, mais le temps d'annotation du corpus d'apprentissage. Par conséquent, l'adaptation des systèmes supervisés à d'autres domaines suppose la création de nouveaux corpus d'apprentissage.

5.3.3. Apprentissage semi-supervisé

C'est dans l'objectif de réduire le coût de créations de ressources, ainsi que d'adapter plus rapidement les systèmes de RCEN à des tâches autres que celles pour lesquelles ils ont été développés, qu'ont été proposées des approches semi-supervisées. Ces approches prennent pour point de départ un ensemble réduit de données annotées, c'est-à-dire d'exemples de chaque classe, à partir desquelles des règles sont induites par généralisation. Ces approches sont compatibles avec les méthodes d'extraction de patron lexico-syntaxiques, comme proposées par M. Hearst [Marti A. Hearst, 1992]. Cette dernière proposait de définir des patrons comme *X such as Y* (par exemple *des animaux comme les chats*) pour récupérer des paires d'hypéronymes et de projeter ensuite ces paires (<*X,Y*>, comme <*animal, chat*>) pour récupérer les séquences de mots qui les séparent et extraire ainsi récursivement d'autres paires.

Cette perspective a été adaptée pour les EN par E. Riloff et R. Jones [Riloff & Jones, 1999] dans une méthode connue sous le nom de « bootstrapping mutuel » : il s'agit d'inférer à la fois des couples (<*EN, Classe*>) et des couples (<*EN, Patron*>). L'algorithme consiste en une boucle qui démarre avec quelques mots-germe de chaque classe (*IBM*, *Toshiba*, etc.) et aucun patron. Les mots-germes (des EN) sont projetés sur un corpus et un score est attribué aux patrons dans lesquels ils ont employés. Le meilleur patron est extrait, ainsi que les nouvelles EN qu'il permet d'obtenir, et la procédure recommence avec ces nouvelles données. Pour fonctionner, un tel algorithme doit répondre à trois questions :

(a) Qu'est-ce qu'un patron ?

(b) Comment mesurer la pertinence d'un patron ?

(c) Quand arrêter la procédure ?

Pour répondre à (a), les auteurs utilisent Autoslog [Riloff, 1993] qui sélectionne des patrons syntaxiques liant un nom et un verbe, ou deux noms (figure 5.6). Ces patrons correspondent à des relations syntaxiques préalablement identifiées par l'analyseur Circus [Lehnert et al., 1991].

PATTERN	EXAMPLE
<subj> passive-verb	<victim> was murdered
<subj> active-verb	<perp> bombed
<subj> verb infin.	<perp> attempted to kill
<subj> aux noun	<victim> was victim
passive-verb <dobj>[1]	killed <victim>
active-verb <dobj>	bombed <target>
infin. <dobj>	to kill <victim>
verb infin. <dobj>	tried to attack <target>
gerund <dobj>	killing <victim>
noun aux <dobj>	fatality was <victim>
noun prep <np>	bomb against <target>
active-verb prep <np>	killed with <instrument>
passive-verb prep <np>	was aimed at <target>

FIG 5.6 – Exemple de couples syntaxiques extraits par Autoslog [Riloff, 1993]

Le score d'un patron (question b) récompense ceux qui permettent d'acquérir le plus d'EN (et non ceux qui permettent d'extraire les plus fréquentes) en se basant sur le rapport entre le nombre d'EN connues F_i et le nombre total de candidats (têtes de syntagmes) obtenus par ce patron N_i :

$$Score(Patron_i) = \frac{F_i}{N_i} . \log_2(F_i)$$

où F_i=nombre d'EN uniques connues et N_i=nombre total de candidats uniques

Étant donné que les patrons récupèrent également des candidats non pertinents, la procédure est rapidement bruitée. Pour y remédier, les auteurs introduisent un calcul supplémentaire pour évaluer la pertinence des candidats extraits. Ces derniers sont évalués par rapport au nombre de patrons dans lesquels ils figurent (d'où la notion de mutualité du bootstrapping) : plus un candidat est extrait par des patrons différents, plus il est susceptible d'appartenir à la catégorie ciblée par ces patrons. Les auteurs sélectionnent les cinq meilleurs mots, et pour ordonner les candidats ayant le même nombre de patrons, ils ajoutent au centième la somme des scores des patrons :

$$Score(candidat_i) = \sum_{k=1}^{N_i} .1 + (0{,}01 . score(Patron_k))$$

Enfin, pour répondre à la question (c), les auteurs ont utilisé des seuils sur le nombre d'itérations (50) et sur le score des patrons extraits.

Une conclusion intéressante de ce travail est que les listes obtenues sont spécifiques à un domaine et qu'elles reflètent la sémantique de ce domaine. Les auteurs remarquent à ce propos que les véhicules et les armes possèdent un nombre de patrons similaires dans les textes terroristes. Ils en concluent qu'il serait logique de considérer qu'ils font partie d'une même catégorie, étant donné que les véhicules sont employés comme des armes dans ce domaine [Riloff & Jones, 1999 : 479]. Remarquons que les auteurs ne proposent pas d'affiner leurs patrons.

Dans une perspective similaire, Collins et Singer [Collins & Singer, 1999] proposent de produire des systèmes semi-supervisés ayant pour apport initial 7 règles : 5, déterminant la classe d'une EN à partir de sa forme et 2, à partir de la forme de son contexte. Les auteurs évaluent leurs

systèmes sur des structures précises dans lesquelles apparaissent les EN : les séquences SN-préposition-SN (où SN est un syntagme nominal) et les séquence SN-SN en apposition, qui sont préalablement repérées par un analyseur syntaxique [Collins, 1996]. Le type (apposition ou préposition) ainsi que la forme (tête syntaxique) de ce contexte sont employés comme traits pour l'apprentissage, à côté de propriétés concernant le mot-cible (capitalisation, présence de points, etc.). Malgré l'absence de connaissances externes (pas de lexiques), les systèmes parviennent à de bons scores pour annoter les entités (Personne, Organisation, Lieu) participant à ces relations locales. Ce travail invite à décomposer en amont la détection des EN en sous-tâches, selon les structures linguistiques auxquelles elles sont associées.

A. Cucchiarelli et P. Velardi [Cucchiarelli & Velardi, 2001] proposent d'utiliser un système de RCEN comme base de départ. Ils combinent ensuite un analyseur syntaxique et des relations de similarité issues de Wordnet pour extraire des relations associées aux EN détectées. Ces relations sont alors employées pour catégoriser des Noms Propres non annotés par le système initial. Les auteurs ne discutent pas de l'intégration du système « bootstrappé » au système initial : il est appliqué aux Npr non détectés.

5.3.4. Choix de l'approche

Cette courte revue des systèmes existants pour la tâche d'EN nous permet de mieux situer notre approche. Notre intérêt va vers les systèmes symboliques car ils sont transparents [Siniakov, 2008 : 31]. En effet, il est possible d'expliquer la cause d'une erreur d'annotation pour les corriger itérativement. À l'inverse, les systèmes supervisés et de manière générale les systèmes statistiques ont l'inconvénient d'être généralement des « boîtes noires » qui ne permettent pas de cibler la cause de l'erreur, ni non plus d'établir un lien entre les indices contextuels et la décision de classification. Les travaux rapportent au mieux l'impact des traits sur les performances : capitalisation, taille de fenêtre, prise en compte d'information syntaxique ou de ressources externes, etc.

La problématique qui nous intéresse est l'adaptation de systèmes symboliques à base de règles. Le problème consiste donc à minimiser le coût d'adaptation d'un système symbolique à de nouvelles entités, domaines, genres, tâches et corpus (cf. *infra* 5.2). En effet, la détection et la correction manuelle d'erreurs pour un système symbolique est une lourde tâche. G. Petasis et ses collègues [Petasis et al., 2001] suggèrent une méthode de détection d'erreurs pour les systèmes symboliques. Elle consiste à entraîner un système supervisé à partir de l'annotation fournie par un système initial à base de règles et d'interpréter leurs désaccords comme indice d'erreur. Les auteurs proposent d'utiliser une matrice de contingence pour comparer les choix communs de classification aux deux systèmes et identifier ainsi les occurrences de divergence. Dans les deux expériences menées, ils ont constaté 95% et 91% d'accord, sans pour autant donner des indications concernant l'exactitude de ces accords : on ne sait pas si les cas d'accord des deux systèmes correspondent ou pas à des erreurs. En analysant les exemples, ils ont pu identifier des absences de lexique, des noms propres ambigus ou encore des règles trop contraintes. Cette méthode de détection d'erreur peut donc être utile mais elle gagnerait à être complétée par une correction automatique.

Les travaux sur le bootstrapping constituent une approche prometteuse pour parvenir à une correction automatique. En effet, les patrons sont extraits à partir du corpus et ils semblent capter des régularités sémantiques (des relations) propres aux classes d'EN ciblées dans ce contexte particulier. Mais une telle approche nécessite des mots-germes ou exemples de départ. Par ailleurs, les travaux de Riloff et Jones sont limités à la création de liste de patrons et d'EN et non à la correction d'un système de RCEN. Or, la greffe d'un module d'extraction de patrons, ainsi qu'une procédure pour les sélectionner (par bootstrapping par exemple), à un système de RCEN peut permettre d'acquérir des règles pour l'adapter automatiquement. Pour y parvenir, il est capital que le

formalisme de représentation des règles du système de RCEN soit identique à celui employé par le module d'extraction. C'est dans cette perspective que nous avons inscrit notre travail, en gardant à l'esprit l'influence exercée par le corpus et le contexte d'application.

Avant de décrire le système que nous avons implémenté pour adapter un système de RCEN symbolique, nous devons présenter le contexte et les travaux préliminaires qui nous y ont conduits. Au risque de nous répéter, nous considérons que les EN sont mieux appréhendées dans le cadre des relations sémantiques auxquelles elles sont associées en contexte (cf. *infra* p145). Ces relations se présentent sous une certaine forme dans le corpus, et font partie des informations que recherche un SQR.

6. Acquisition de cadre d'EI

Notre recherche sur les EN s'est effectuée dans le cadre d'un projet de SQR, le système RITEL, qui fait usage d'un analyseur linguistique Ritel-nca, incorporant la tâche d'EN. Nous présenterons tout d'abord le projet et le système symbolique à l'œuvre. Dans un second temps, nous présenterons l'approche collocationnelle développée pour l'extraction de cadres d'EI (Extraction d'Information), permettant à la fois la désambiguïsation et l'extraction de relations. Nous présenterons ensuite une amélioration du système qui repose sur l'ajout d'une couche analyse syntaxique en constituants, avant de terminer sur un cas particulier, l'extraction de citations.

6.1. Contexte de recherche

RITEL [Rosset et al., 2005] est un SQR interactif à domaine ouvert dont l'objectif est de permettre à un utilisateur de dialoguer avec un système de recherche d'information généraliste. Développé au LIMSI, il bénéficie de l'expérience acquise dans le développement des systèmes qui l'ont précédé (Arise, RailTel, Mask[37]). Historiquement, les applications ont progressivement évolué, de l'information sur le domaine aéronautique [Bennacef et al., 1994], la réservation de billets de trains [Lamel et al., 2000] à des questions à domaine ouvert. Les sources ont également évolué, des bases de données structurées, à des textes non structurés d'information générale.

6.1.1. Le système RITEL

Le système RITEL se décline globalement en 5 types de traitements et est schématisé dans la figure (6.1) :

- Reconnaissance de la parole
- Analyse linguistique
- Gestion du dialogue et/ou Recherche d'information
- Génération automatique
- Synthèse de la parole

L'analyse linguistique Ritel-nca (Rnc), ou analyse non-contextuelle (parce qu'elle ne tient pas compte du contexte de dialogue pour prendre des décisions), permet d'annoter les transcriptions issues de la reconnaissance vocale. Comme on peut l'observer sur le schéma, c'est ce même module qui analyse les documents dans lesquels s'effectue la RI. Cela signifie que ce module prend à la fois en charge l'écrit et l'oral (les corpus peuvent être constitués d'articles de presse comme de transcriptions radiophoniques). C'est un choix stratégique délibéré : il permet d'utiliser la même représentation linguistique dans les questions et dans les réponses, autant à l'oral qu'à l'écrit.

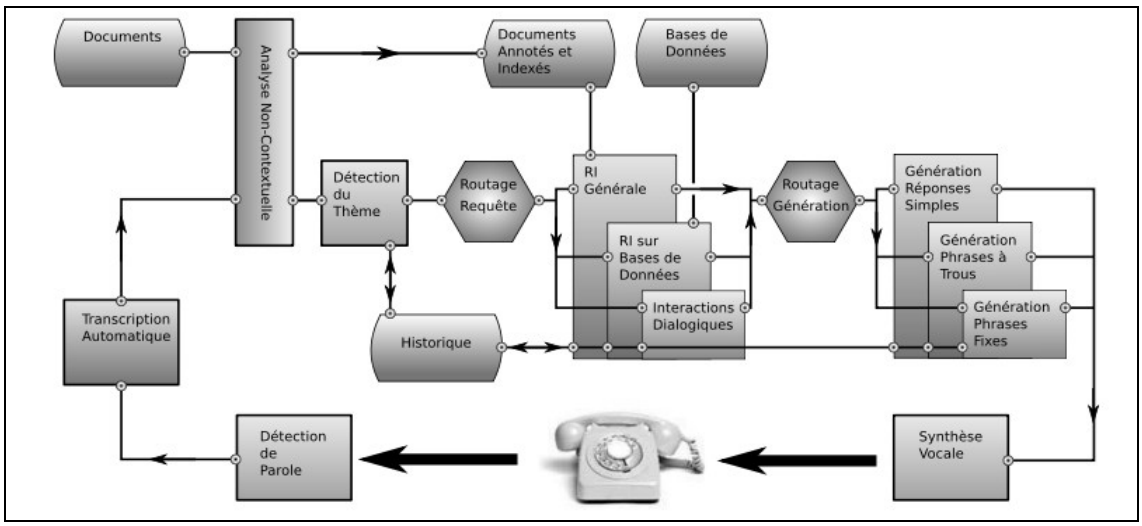


FIG 6.1 – Architecture du système RITEL [Rosset et al., 2005 : 159]

37 http://www.limsi.fr/tlp/projects.html

- Ritel-nca est un système symbolique à base de règles qui repose sur un moteur d'expressions régulières dédié, Wmatch [Galibert, 2009]. Il utilise un grand nombre de lexiques pour définir des catégories (appelées entités-R[38], au nombre de 300) de nature variable (inspiré en partie de la taxonomie de Sekine, [Sekine & Nobata, 2004] ; tableau 6.1) à partir de règles.

Entités nommées	<_org> NIST </> <_eve> festival de Cannes de 2006 </> qui a dit <_cit> veni vidi vici </>
Entités non précises	<_Eve> festival de Cannes </> le <_Pers> president </> a déclaré ...
Entités étendues multi-niveaux	Fonctions, titres (président, professeur, évêque...) couleurs, animaux...
Super classes hiérarchiques	évêque → fonction religieuse → fonction
Marqueurs thématiques	Je m' intéresse aux <_litterature> romans </> de ... qui a gagné le <_sport> Mondial </> de 1998
Marqueurs interrogatifs	<_Qqui> qui </< a écrit ce livre <_Qmesure> de combien </> d' heures dure ...
Marqueurs d'interaction	<_DA_close> au-revoir </> <_DA_yes> oui s'il vous plaît </>
Mots composés	les <_NN> logiciels de base de données </> sont ... les <_NN> élections multi-raciales </>
Chunks verbaux	il a <_action> gagné</> ... ils <_action> prendront part à </> ...
Entités linguistiques	<_adj_comp> le plus gros </> exportateur... cela se produit <_adv> souvent </> quand ...

Tableau 6.1 – Types d'Entités-R de la grammaire de Ritel-nca [Rosset et al., 2005 : 163]

Les règles définissent des grammaires régulières dédiées à la détection d'un type d'entité-R, elles sont donc relativement proches des grammaires sémantiques ([Burton, 1977] ; [Gavaldà, 2000]) : les règles s'appliquent à des catégories sémantiques. De nombreuses passes sont effectuées pour appliquer ces règles et des profils peuvent être définis pour choisir et ordonnancer leur application. Les entités-R peuvent ainsi être hiérarchisées, imbriquées et corrigées par des règles de plus haut niveau et les textes sont annotés par des structures arborescentes (exemple 193, illustré en figure 6.2). La f-mesure associée à la classification d'entités nommées (Organisation, Personne, Lieu) est de 0,8 sur l'écrit et à hauteur de l'état de l'art pour les corpus oraux [Rosset et al., 2005 : 167].

(193)Patricia Highsmith est morte le 4 février 1995 dans un hôpital de Locarno.

38 Pour éviter la confusion entre entités nommées et entités Ritel, nous adopterons la forme *entité-R* pour ces derniers. Les entités-R désignent les étiquettes du système Ritel et couvrent les EN mais également d'autres catégories, comme les tags morpho-syntaxiques. Elles sont individuellement signalées dans le texte précédées un tiret bas comme *_pers*. On trouvera en annexe la liste des entités-R utilisées dans notre système.

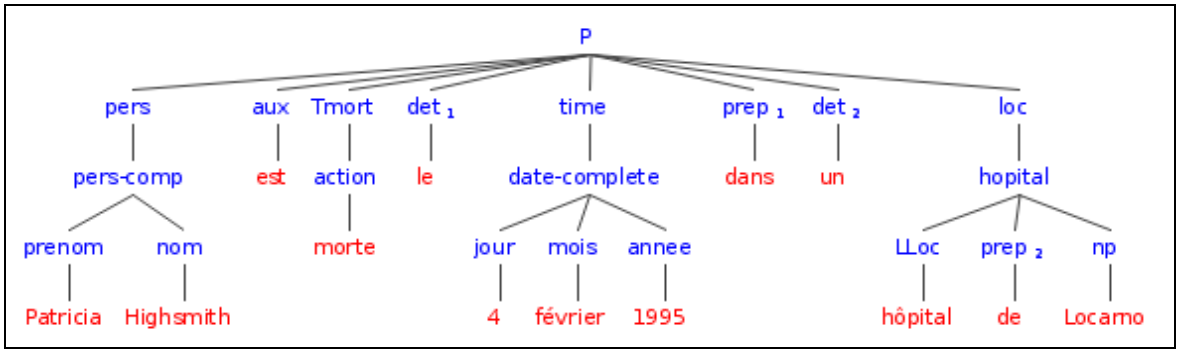


FIG 6.2 – Exemple de sortie d'annotation de Ritel-nca (exemple 193)

- La détection du thème précède la phase de RI : son objectif est de détecter et de stocker les unités relevant du même thème dans un historique. Lorsque ce module détecte un changement de thèmes (à partir de marqueurs de l'analyse linguistique), l'historique est ignorée dans la RI.
- Le module de Routage de la requête envoie l'énoncé vers le module de RI approprié : RI sur base de données, Interaction dialogique ou RI générale. Nous nous intéresserons au dernier.
- La RI Générale nécessite l'analyse de l'énoncé utilisateur, conçu comme l'analyse de la question d'un SQR. Celle-ci est formalisée par un Descripteur de Recherche (DDR) qui représente les éléments pertinents de l'énoncé à retrouver dans l'environnement de la réponse, ainsi que leurs expansions pondérées (lemme, synonymes, ...) et le type de la question [Galibert, 2009 : 95–99]. Par exemple, le DDR correspondant à l'exemple (194) est décrit dans la figure (6.3).

  (194)Quand a été assassiné Hans Krasa ?

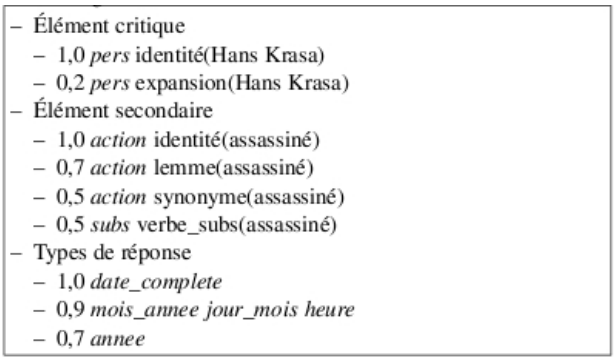


FIG 6.3 – DDR obtenu à partir de l'exemple 194 [*ibid.* : 96]

Le type de réponse attendu est bien une date, l'élément devant absolument figurer dans le contexte est le Npr de type Personne *Hans Krasa* ainsi que le verbe (ou *action*) *assassiner* (en y ajoutant ses variantes sémantiques et morpho-syntaxiques). L'intérêt de ce DDR est d'être à la fois lisible pour le concepteur et structuré pour lancer automatiquement une requête. La requête ainsi définie est suivie des trois étapes classiques de recherche d'information dans un SQR : la sélection des documents, la sélection de passages et l'extraction de la réponse. Pour plus de précision sur l'originalité de ce système, nous renvoyons le lecteur à la thèse de Galibert [*ibid.* : 101–114].

- Une fois la réponse obtenue (dans le cas où effectivement il y en a une), le système génère une réponse (ou plusieurs), qui est ensuite synthétisée vocalement. Plusieurs stratégies de génération sont possibles, de la génération de patrons de réponse complexes, jusqu'à la demande de précisions auprès de l'utilisateur [Rosset et al., 2005 : 173–175].

Le système RITEL est complexe mais des efforts ont été entrepris pour retracer les erreurs et faciliter leur correction (grâce notamment au DDR et au format de représentation unique fourni par l'analyseur linguistique). Comme on a pu le voir, l'analyseur linguistique joue un rôle important : c'est en partie sur lui que reposent la détection de thème, la définition du DDR, l'indexation et l'extraction de la réponse. Les évaluations détaillées de RITEL ont montré que le type de la réponse était source de nombreuses erreurs [Galibert, 2009 : 152][39]. D'après O. Galibert, l'amélioration principale repose sur l'addition d'informations linguistiques[40], pouvant être exploitées par le SQR :

> « Mais la principale faiblesse est la limite sur ce que l'analyse peut représenter. Comme nous l'avons vu dans la première partie la représentation ne permet pas d'annoter les relations à longue distance. Or nous pensons que des relations sémantiques de qualité liant les éléments trouvés par l'analyse permettraient d'obtenir une bien meilleure qualité au niveau du score de réponse. Cependant, représentation mise à part, poser de telles relations de façon fiable semble très difficile sans de grandes ressources linguistiques. Concevoir un moteur de règles capable de travailler sur de telles relations, écrire des règles pour les établir d'une manière fiable et les exploiter ensuite dans un système Question-Réponse restent des problèmes ouverts. » [Galibert, 2009 : 164]

Cette observation constitue le point de départ de nos expériences sur les EN, sachant notre approche. Avant de proposer des réponses sur le problème de la représentation à adopter, nous avons étudié des corpus annotés par le système Ritel-nca pour évaluer à la fois le problème de l'extraction de relations sémantiques et de la catégorisation des Npr. L'approche définie consiste à exploiter les technologies de TAL pour réaliser une analyse de corpus dans une perspective spécifique d'extraction d'information (chapitre 5). Nous avons ciblé notre étude sur les personnes.

6.1.2. Corpus d'étude

Dans ce projet, nous avons utilisé deux corpus de taille et de nature différentes : un corpus de biographie et un corpus de presse.

Le premier corpus est composé de 430 biographies extraites automatiquement de l'encyclopédie Wikipedia[41]. La longueur moyenne des textes est de 1700 mots additionnant un total de 727 000 formes pour le corpus. Wikipédia est une encyclopédie multilingue[42], libre et collaborative et la qualité de ses articles a fait l'objet de nombreuses controverses. De nombreux travaux proposent des méthodes quantitatives pour mesurer la qualité d'un article ou leur contenu [Kittur & Kraut, 2008] ;[Liu & Ram, 2011], un certain nombre de statistiques peuvent également être obtenues sur le site-même de l'encyclopédie. En TAL, c'est devenue une source de connaissance privilégiée, de par sa disponibilité, son actualité et sa structuration très riche. Elle a notamment été

39 Ritel a participé aux trois éditions de Qast (campagne dédiée aux-questions-réponses sur de l'oral), sur l'anglais, l'espagnol et le français ainsi qu'à la campagne Quæro sur l'anglais et le français. Pour les résultats, voir [Galibert, 2009 : 137–145].

40 Une autre voie a été explorée par G. Bernard dans le cadre de sa thèse sur RITEL : l'amélioration du réordonnancement de réponse [Bernard, 2011].

41 http://www.wikipedia.org/

42 Wikipedia compte plus d'un million d'article et est le 7e site le plus consulté en France (2010).

utilisée comme ressource pour la classification d'EN [Bunescu & Pasca, 2006]. Chaque article est en effet associé à des catégories qui facilitent l'extraction de ceux qui concernent un type d'EN recherchée (personne, organisation, etc.). Par exemple, pour la biographie d'Abdou Diouf, les catégories associées à sa biographie (illustrée en 6.4) sont listées en (195).

(195)*Catégories :* Personnalité du Parti socialiste (Sénégal), Premier ministre du Sénégal, Président du Sénégal, Étudiant de l'université Cheikh Anta Diop, Docteur honoris causa, Francophonie, Naissance en 1935, Naissance à Louga

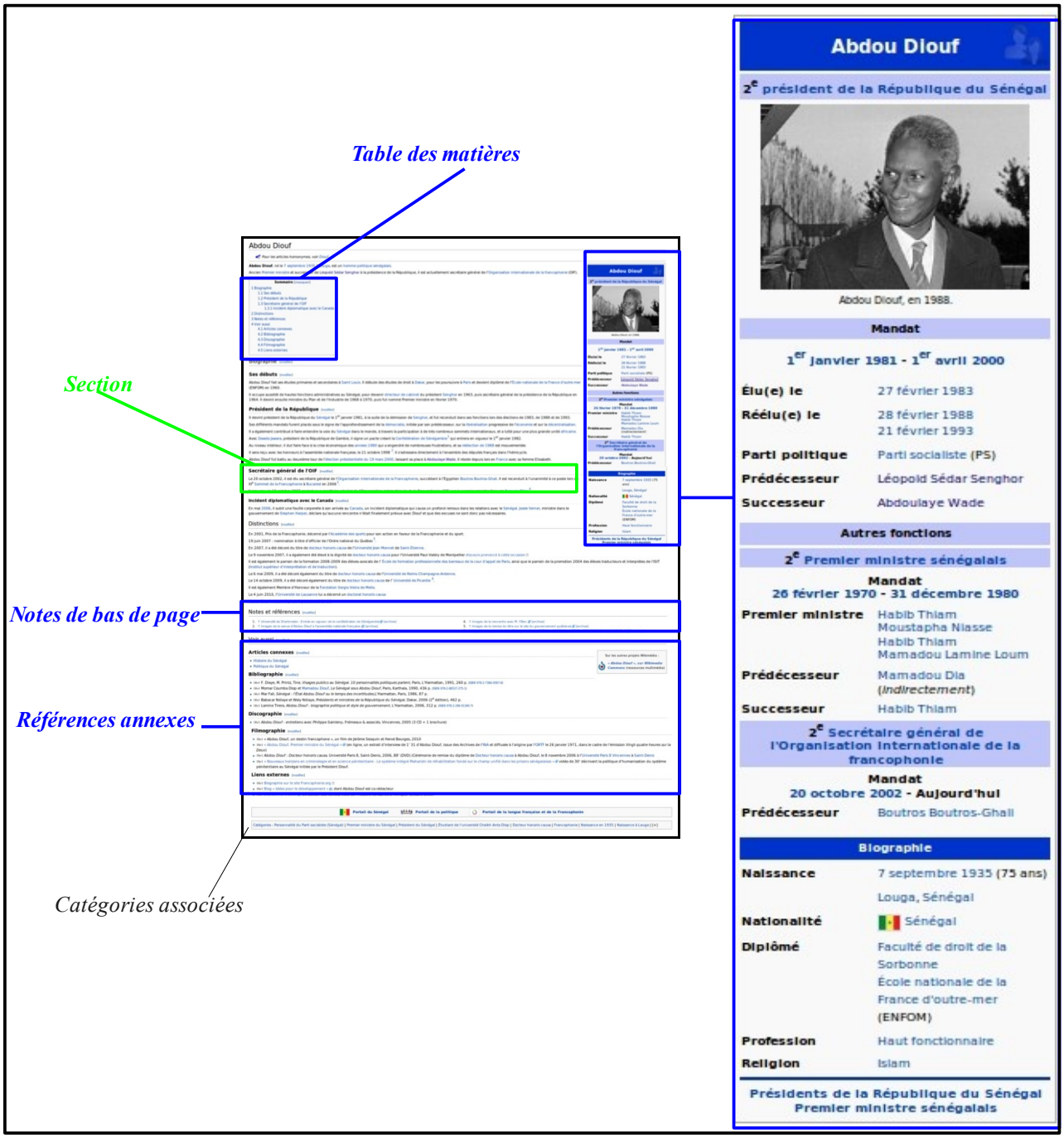


FIG 6.4 – Exemple de page biographique et Infobox associée (Wikipedia)

Les catégories sont organisées dans une arborescence dont les pages sont les feuilles, ce qui permet de retrouver des pages similaires en fonction d'un critère donné (par exemple, tous les *premiers ministres du Sénégal*). Lorsque le sujet est suffisamment important, il donne lieu à la naissance de portails comme celui des biographies. On peut ainsi filtrer les biographies en fonction des périodes, du métier, de la nationalité, ou encore par cause de mort. Certaines personnes peuvent même avoir une catégorie associée, ou plutôt, certains Npr peuvent être des catégories. La catégorie *Charles de Gaulle* contient par exemple un article principal, 3 sous-catégories et 43 pages associées à cette catégorie (figure 6.5 ; cf. *infra* p141).

FIG 6.5 – Pages de la catégorie Charles de Gaulle

Les pages contiennent divers types d'information structurée (les infoboxs par exemple), semi-structurée (tableaux, listes à puces, titres de section, etc.) ou non structurée (texte). Les personnalités sont souvent associées à des infoboxs (figure 6.4), tableaux microfiche qui résument les informations majeures en couple <*attribut, valeur*>. Dans les références annexes, on peut trouver une bibliographie sur/de l'auteur, encodée dans un tableau au format relativement fixe (conventions d'écriture de l'encyclopédie). Malgré des conventions de normalisation strictes (Wikipedia possède son propre langage de mise en forme), l'analyse des fichiers XML des pages Wikipedia montre que les schémas d'écriture sont très variables (lorsqu'il ne s'agit pas d'erreurs), ce qui pose des problèmes pour l'extraction automatique du texte à partir des dumps (version datée de l'encyclopédie).

Le second corpus dont nous disposons est l'ensemble des articles de presse du journal LeMonde de l'année 2004 (22 millions de formes). Ce corpus est bien moins homogène que le corpus de biographies, car il traite de nombreux sujets (finance, résultats sportifs, etc.). Les articles sont associés à des catégories, qui correspondent à des rubriques (Sport, International, Culture, Économie, Politique, etc.) indiquées par différentes balises XML ; il serait donc possible de les exploiter pour le diviser en tranches plus homogènes (en sections). Néanmoins notre but principal était de tester notre approche sur un corpus de plusieurs millions de mots, et qui soit relativement général. Le type et la taille d'articles sont également variables (éditorial, brève, critique, analyse, etc.). Parmi les 43 450 articles du corpus, le plus long reporte les résultats des élections européennes en France et comporte plus de 15 000 formes. Nous avons travaillé sur plusieurs versions de ce corpus, notamment sur un corpus de développement constituant les trois quarts du corpus, soit 32750 articles pour 17 millions de formes. Un quart était réservé pour évaluer notre système décrit

chapitre 10. Ce corpus a été annoté dans les mêmes conditions que le corpus de biographies, mais nous avons dû réaliser un certain nombre de prétraitements génériques qui puissent s'appliquer à tout type de corpus. À cette fin, nous avons réalisé une chaîne de traitement incorporant des modules pour l'extraction du texte, l'encodage, la segmentation en mots, la normalisation, l'étiquetage, etc.

6.2. Extraction d'Information biographique

Les biographies constituent un terrain d'étude privilégié pour les EN de personne. On y retrouve de nombreuses occurrences du référent en question (niveau d'identité référentielle ; cf. *infra* 1.3.3) et chacune de ces occurrences peut être associée à une information qui permet de le classer (niveau de classification référentielle ; cf. *infra* 1.3.3). Par exemple, on y trouve les dates marquantes de sa vie ou les différentes professions exercées. L'application générale est la population de bases de données (les informations qui figurent dans les infoboxs par exemple, cf. *infra* figure 6.4) [Garera & Yarowsky, 2009]. Ces relations liées aux personnes ont été définies pour être exploitées par un SQR, dans le cas où la question attend une de ces informations.

6.2.1. Analyse collocationnelle

L'analyse des catégories sémantiques majeures gravitant autour des personnes dans ce corpus à partir des entités-R est la première étape pour identifier des relations. Les tableaux (6.2) et (6.3) indiquent les 30 entités-R (de plus haut niveau dans la hiérarchie du système Rnc) et formes les plus fréquentes dans le corpus de biographie, leurs proportions par rapport au nombre total d'occurrences du corpus ainsi que les fréquences et proportions cumulées.

Entité	Fréq.	Fréq. Cumul.	Prop.	Prop. Cumul.
<_punct>	102850	102850	14,58%	14,58%
<_prep>	88840	191690	12,59%	27,17%
<_det>	83551	275241	11,84%	39,01%
<_subs>	55328	330569	7,84%	46,85%
<_action>	55024	385593	7,80%	54,65%
<_pers>	34536	420129	4,89%	59,55%
<_pronom>	33791	453920	4,79%	64,33%
<_NN>	25125	479045	3,56%	67,90%
<_conjc>	20320	499365	2,88%	70,78%
<_adjectif>	20061	519426	2,84%	73,62%
<_mot_inconnu>	18808	538234	2,67%	76,28%
<_adv>	18318	556552	2,60%	78,88%
<_aux>	16989	573541	2,41%	81,29%
<1>	13326	586867	1,89%	83,18%
<_np>	11573	598440	1,64%	84,82%
<_time>	11099	609539	1,57%	86,39%
<_conjs>	10473	620012	1,48%	87,87%
<_loc>	9351	629363	1,33%	89,20%
<_loc_adv>	7427	636790	1,05%	90,25%
<_DDnon>	4093	640883	0,58%	90,83%
<_locorg>	3959	644842	0,56%	91,39%
<_gerondif>	3834	648676	0,54%	91,94%
<_pers_act>	3466	652142	0,49%	92,43%
<_range_objet>	3173	655315	0,45%	92,88%
<_pval>	2363	657678	0,33%	93,21%
<_org>	2184	659862	0,31%	93,52%
<_measure_phys>	2164	662026	0,31%	93,83%
<_Ffamille>	2043	664069	0,29%	94,12%
<_Organisation>	1984	666053	0,28%	94,40%
<_fonctions>	1842	667895	0,26%	94,66%

Tableau 6.2 – Statistiques d'entités-R dans les biographies

Forme	Fréq.	Fréq. Cumul.	Prop.	Prop. Cumul.
,	42094	42094	5,97%	5,97%
.	26303	68397	3,73%	9,69%
de	25864	94261	3,67%	13,36%
la	17159	111420	2,43%	15,79%
"	16690	128110	2,37%	18,16%
et	15068	143178	2,14%	20,29%
le	14652	157830	2,08%	22,37%
l'	13650	171480	1,93%	24,30%
à	12281	183761	1,74%	26,04%
il	11485	195246	1,63%	27,67%
en	10328	205574	1,46%	29,13%
les	8420	213994	1,19%	30,33%
un	7031	221025	1,00%	31,32%
est	6886	227911	0,98%	32,30%
d'	6780	234691	0,96%	33,26%
une	6290	240981	0,89%	34,15%
des	6159	247140	0,87%	35,03%
(	6015	253155	0,85%	35,88%
)	5990	259145	0,85%	36,73%
dans	5630	264775	0,80%	37,52%
du	5332	270107	0,76%	38,28%
son	5188	275295	0,74%	39,02%
qui	4883	280178	0,69%	39,71%
pour	4424	284602	0,63%	40,33%
par	4100	288702	0,58%	40,92%
:	3648	292350	0,52%	41,43%
au	3633	295983	0,51%	41,95%
sa	3548	299531	0,50%	42,45%
que	3385	302916	0,48%	42,93%
a	3114	306030	0,44%	43,37%

Tableau 6.3 – Statistiques des formes dans les biographies

Dans ce corpus, on observe que 90% des occurrences d'entités-R sont couvertes par 18 unités, alors que les 30 formes les plus fréquentes couvrent moins de 44% des occurrences : les entités-R réduisent la variation, bien qu'on ne sache pas si les regroupements effectués soient pertinents[43].

43 Parmi les entités-R, on trouve les catégories morpho-syntaxiques : *_subs*, *_prep*, *_det*, *_action*, *_pronom*, *_conjc*,

Smadja [Smadja, 1993] propose d'utiliser l'environnement d'un mot pour construire un profil contextuel. L'environnement est quadrillé sous forme de tableau, dont les colonnes correspondent aux positions relatives vis-à-vis du nœud-cible (limitées à 5 mots avant et après) et les lignes, aux unités apparaissant dans au moins une de ces positions. Nous avons appliqué cette méthode pour déterminer le profil de la catégorie personne, en combinant les niveaux de forme ou d'entité-R.

Le profil contextuel de l'entité-R *_pers* est composé de 240 entités-R dont 143 ont une fréquence en contexte supérieure ou égale à 5. La large gamme d'entités-R s'explique notamment par le fait que les nombres non typés ou non rattachés sont considérés comme des entités-R (*<8>* par exemple). Ce profil peut être ordonné en fonction de la fréquence dans une position donnée ou en fonction de la fréquence globale en contexte et faire ainsi émerger des préférences positionnelles de certaines entités-R. Le tableau (6.4) présente les probabilités d'occurrence des « collocats d'entités-R » triés en fonction de la position L1 (L pour gauche, R pour droite).

Contexte	Fréq.(y)	Fréq. (x,y)	L	R	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
<_Qpourquoi>	70	9	6	3	0	0,11	0	0	0,56	0,11	0,11	0	0,11	0
<_annonce_pers>	150	126	86	35	0,04	0,02	0,06	0,04	0,56	0,08	0,08	0,02	0,02	0,07
<_annonce_titre>	44	63	46	16	0,02	0,05	0,11	0,05	0,52	0	0,13	0,03	0,06	0,03
<_insrt>	256	578	360	178	0,07	0,07	0,06	0,04	0,44	0,06	0,08	0,05	0,06	0,06
<_Qorg>	29	7	4	2	0,14	0	0,14	0	0,43	0	0,14	0	0,14	0
<_pers_fonct>	1580	1570	903	551	0,07	0,05	0,06	0,08	0,38	0,04	0,08	0,11	0,06	0,06
<00>	41	16	13	2	0,06	0,19	0,13	0,13	0,38	0	0	0	0,06	0,06
<_Pers>	308	160	103	43	0,09	0,11	0,11	0,08	0,34	0	0,06	0,08	0,06	0,07
<_Qqdef>	20	10	8	2	0	0,2	0,2	0,1	0,3	0	0,1	0	0	0,1
<9>	27	11	5	5	0,09	0	0	0,18	0,27	0	0,09	0,18	0	0,18
<_conjs>	10473	3435	1593	1466	0,11	0,09	0,1	0,05	0,22	0,06	0,11	0,1	0,08	0,08
<_adv_compos>	1160	469	249	158	0,13	0,16	0,1	0,04	0,22	0,05	0,07	0,06	0,08	0,07
<_titre>	330	295	186	93	0,05	0,06	0,06	0,28	0,22	0	0,08	0,08	0,07	0,07
<_Tlangue>	56	9	7	1	0,11	0	0,22	0,33	0,22	0	0	0	0,11	0
<_conjs_compos>	1684	545	239	261	0,08	0,1	0,1	0,03	0,21	0,14	0,12	0,08	0,06	0,07
<_prep>	88840	35316	17452	14378	0,1	0,12	0,11	0,05	0,21	0,07	0,07	0,09	0,09	0,09
<_org_pol>	339	135	82	31	0,16	0,1	0,14	0,18	0,19	0,01	0,04	0,04	0,06	0,07
<_locorg>	3959	1902	1027	664	0,11	0,09	0,1	0,15	0,19	0	0,1	0,08	0,07	0,1
<_Monnaie>	124	244	93	113	0,16	0,02	0,16	0,01	0,18	0,15	0,01	0,14	0,02	0,14
<_Qdate>	49	11	5	5	0,09	0,09	0,09	0,09	0,18	0,09	0,18	0	0	0,18
<_Qou>	1176	387	137	204	0,12	0,09	0,09	0,01	0,17	0,11	0,18	0,09	0,09	0,07
<_conjc>	20320	8780	3290	4733	0,09	0,08	0,07	0,06	0,16	0,21	0,08	0,08	0,08	0,09
<_punct>	102850	53493	21906	26516	0,09	0,09	0,09	0,07	0,16	0,21	0,05	0,09	0,07	0,08
<_orig>	1618	538	297	150	0,17	0,1	0,12	0,17	0,16	0,04	0,01	0,04	0,08	0,11
<_Prix>	127	49	19	28	0,04	0,06	0,04	0,12	0,16	0	0,04	0,16	0,16	0,2
<_culte>	32	14	5	6	0,21	0,07	0	0,14	0,14	0	0	0,07	0,14	0,21
<_musee>	61	23	13	8	0,09	0,22	0,13	0,09	0,13	0,04	0,04	0	0,17	0,09
<_U_info>	1307	293	162	67	0,22	0,17	0,11	0,14	0,13	0,02	0,09	0,04	0,03	0,05
<_date_relative>	18	8	7	0	0,13	0,13	0	0,63	0,13	0	0	0	0	0
<_DDouverture>	25	8	4	3	0,13	0,13	0	0,25	0,13	0	0	0	0,13	0,25

Tableau 6.4 – Profil contextuel de l'entité-R _pers dans le corpus de biographie

_adjectif, *_adv*, *_aux*, *<1>* (*1, une* ou *un*), *_conjs*, *_loc_adv* (locutions adverbiales), *_DDnon* (négation), *_gerondif*, et *_range_objet* (formes comparatives et superlatives). La ponctuation (*_punct*) est l'entité-R la plus fréquente. L'entité-R *_NN* désigne des composés nominaux dont une partie est extraite du dictionnaire DELAC (cf. http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/delac.html). Les mots inconnus peuvent être dus à des erreurs lors de l'extraction du texte de la biographie, à des bugs d'encodage, ou à des formes absentes des lexiques. *_np* désigne des noms propres non typés. Le reste est composé de catégories sémantiques dont *_pers* (Personne) est la plus fréquente : _time (temporel), _loc (lieu), *_locorg* (lieu ou organisation), *_pers_act* (activité humaine), *_pval* (valeur numérique faible), _org (organisation), *_measure_phys* (mesure physique), *_Ffamille* (parenté), *_Organisation* (nom générique d'organisation) et *_fonctions* (fonction humaine).

Le tableau illustre un des problèmes liés au calcul des collocats dans une fenêtre de mots : la fréquence en contexte *Fréq.(x, y)* peut surpasser la fréquence réelle *Fréq.(y)* (comme pour *_annonce_titre* ou *_insrt*). En effet, dans le cas où plusieurs entités-R *_pers* sont textuellement proches, les éléments qui apparaissent dans leur environnement seront comptés plusieurs fois. On peut néanmoins se faire une idée de l'importance d'une position pour une entité-R donnée en rapportant la fréquence de la position à la fréquence en contexte, comme nous l'avons fait. Cette « préférence positionnelle » peut être interprétée comme un indice de figement ou d'association significative. Par exemple, lorsque l'entité-R *_annonce_pers* (*madame*, *monsieur*, *docteur*, etc.) apparaît dans l'environnement d'une personne, elle est préférentiellement à sa gauche (0,56).
On peut classer les entités-R du tableau (6.4) en trois groupes de données.

- Les éléments à faible fréquence sont souvent des relateurs : des pronoms interrogatifs (majuscule *Q* initiale) typés par entité-R (*Qdate*, *QOrg*, etc.) ou par question (*Qpourquoi*) ; on trouve également des conjonctions de subordination simples (*_conjs*) ou composés (*_conjs_compos*) et les prépositions (*_prep*).
- Le deuxième groupe correspond à un ensemble de déclencheurs de la catégorie Personne ; on y trouve les initiales, titres et diverses annonces de personne (*_annonce_titre*, *_annonce_pers*, *_insrt*, *_titre*) qui ne sont pas considérés comme partie de la dénomination d'une personne (conventions des campagnes MUC), mais également des noms communs d'humain (*_Pers*) comme *homme*, *personne*, *héros* et des noms de fonctions *_pers_fonct* comme *tsar*, *sénateur* ou *pape*.
- Le troisième groupe est constitué d'entités-R ou déclencheurs sémantiquement différents de *_pers* et sont des indicateurs d'erreur de normalisation ou d'incomplétude des règles. Par exemple, la forme *mark* est systématiquement annoté en *_Monnaie* même dans la séquence *mark Twain* ; la présence de l'entité-R *_langue* est due à l'ambiguïté de noms de nationalité (*_orig*) comme *croate* ; on trouve encore les entités-R *_Prix* comme *prix Charles Cros* (où *Cros* n'a pas été rattaché) ou *_culte* comme *Église de Sutton Courtenay* (où *Courtenay* n'a pas été rattaché) qui illustrent les difficultés de détection d'EN. Lorsqu'un lieu ou une organisation apparaît immédiatement à gauche, il peut également correspondre à une erreur.

Linguistiquement donc, il semble que les éléments que l'on trouve préférentiellement à gauche d'une personne fassent partie du même syntagme que l'EN ou constituent des frontières de ce syntagme.

En classant les collocats en fonction de la position R1, les préférences sont moins fortes : on rencontre majoritairement des auxiliaires (0,35), des modaux (0,26) et des verbes ou sous-classes de verbes (localisation, citation, naissance et mort). On peut établir les mêmes types de profils en se basant sur les formes.

Nous avons choisi de réduire l'analyse en nous focalisant sur une entité-R-pivot, susceptible d'associer les personnes comme élément central d'un cadre d'information. Nous nous sommes concentré sur l'entité-R *_Tmort* qui lorsqu'elle est en relation avec une personne, peut également être accompagnée d'autres informations comme la date ou le lieu de la mort. L'entité-R *_Tmort* ne regroupe pas uniquement les formes du verbe *mourir* et *décéder*, elle contient également leurs nominalisations (tableau 6.5).

Forme	Fréq.
mort	523 (66%)
décés	76 (9.6%)
mourut	58 (7.3%)
décédé	37 (4.6%)
décède	37 (4.6%)
morte	21 (2.6%)
décédée	17 (2.1%)
morts	5
décédés	4
mortes	4
décéda	4
moururent	3
décédées	1
décèdent	1

Tableau 6.5 – Formes associées à l'entité-R _Tmort dans le corpus de biographie

Le profil partiel de cette entité-R *Tmort* est illustré dans le tableau (6.6) trié selon la fréquence en contexte *Fréq. (x,y)*.

Contexte	Fréq.(y)	Fréq.(x,y)	L	R	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
<_punct>	102850	1239	498	620	0,1	0,11	0,11	0,11	0,07	0,15	0,02	0,13	0,1	0,1
<_prep>	88840	1232	488	631	0,09	0,08	0,09	0,21	0,01	0,25	0,04	0,09	0,08	0,05
<_det>	83551	1168	560	495	0,1	0,08	0,06	0,03	0,31	0,14	0,1	0,06	0,06	0,07
<_time>	11099	342	123	197	0,06	0,12	0,09	0,15	0	0	0,38	0,04	0,09	0,06
<_action>	55024	320	137	142	0,13	0,18	0,13	0,11	0,02	0,03	0,05	0,14	0,1	0,13
<_subs>	55328	310	116	154	0,13	0,14	0,17	0,05	0,01	0	0,08	0,19	0,1	0,13
<_pers>	34536	300	112	148	0,13	0,08	0,09	0,13	0,08	0,03	0,25	0,07	0,09	0,06
<_conjc>	20320	281	180	80	0,07	0,06	0,05	0,09	0,44	0,08	0,02	0,05	0,07	0,06
<_loc>	9351	256	121	101	0,13	0,13	0,18	0,13	0,02	0	0,14	0,06	0,14	0,05
<_pronom>	33791	220	100	85	0,16	0,08	0,06	0,12	0,19	0,05	0,12	0,04	0,09	0,09
<_aux>	16989	193	89	81	0,12	0,08	0,06	0,04	0,29	0,06	0,06	0,05	0,15	0,11

Tableau 6.6 – Profil contextuel de l'entité-R _Tmort

Si la ponctuation apparaît fréquemment dans le contexte de cette entité-R, sa distribution dans les différentes positions est relativement homogène (plus faible écart-type de 0,03). En revanche, on observe que les prépositions apparaissent préférentiellement en position R1 et L2 (écart-type de 0,07). On remarque également que les entités-R *_time*, *_pers* et *_loc* ont tendance à éviter les positions environnant l'entité-R *_Tmort*. On peut faire l'hypothèse que leur absence en position R1 et L1 s'explique par le fait que, en contexte, cette position est occupée par une préposition qui permet de les lier au cadre.

Si on se concentre sur la position R1, on constate que la séquences d'entités-R [_Tmort _prep] couvre 40% des occurrences de l'entité-R *_Tmort*. Ces occurrences se déclinent en fonction des formes de préposition illustrées dans le tableau (6.7).

Patron	Fréq.	Prop.
de <_pers>	45	0,15
en <_time>	42	0,14
de <_det>	36	0,12
à <_loc>	34	0,11
d' <1>	26	0,08
à <_det>	14	0,05
de <_locorg>	10	0,03
d' <_pers>	8	0,03
dans <_det>	5	0,02
pendant <_det>	5	0,02
pour <_det>	5	0,02
dans <1>	5	0,02
de <_subs>	4	0,01
de <_mot_inconnu>	4	0,01
après <_det>	3	0,01
au <_subs>	3	0,01
à <_locorg>	3	0,01
en <_subs>	3	0,01
d' <_adv>	3	0,01
à <_annee_dur>	3	0,01

Tableau 6.7 – Séquences en position R1 et R2 de l'entité-R _Tmort

On retrouve effectivement ces entités-R, chacune étant différenciée par une préposition, comme *de* pour *_pers*, *en* pour *_time* et *à* pour *_loc*. On peut donc utiliser cette représentation pour définir des règles d'extraction de relation ou patrons avec Tmort. Nous avons distingué les patrons en fonction du type de relation binaire exprimée. Par exemple la relation *mort(personne)* peut être identifiée au moyen du patron *[_Tmort _prep:de _pers]*. Les patrons peuvent être représentés sous forme d'automates à états finis qui manipulent les deux niveaux de représentation et le critère de position. Pour chaque automate, on peut extraire l'argument de la relation qu'il reconnaît (figure 6.6).

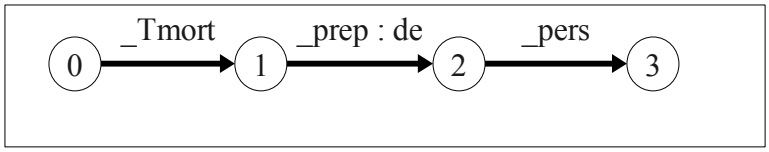


FIG 6.6 – Automate de reconnaissance de la relation Mort(Personne)

La position R1 n'est que le point de départ : comme on l'observe dans le tableau (6.7), certains patrons comme *[_Tmort de _det]* sont incomplets et nécessitent que les positions plus à droite soient explorées. Nous avons exploré l'espace à droite de la totalité des occurrences de la séquence *[_Tmort _prep]*, jusqu'à identification d'une relation. Nous avons ainsi défini 81 patrons qui se répartissent en 5 relations (tableau 6.8).

Relations	Occurrences	patrons
mort(personne)	116	25
mort(date)	75	27
mort(lieu)	56	15
mort(cause)	37	10
mort(âge)	11	4

Tableau 6.8 – Relations associées à l'entité-R _Tmort

On observe que la couverture d'une relation ne croît qu'aux dépens d'une augmentation du nombre de patrons : il faut par exemple 25 patrons différents pour caractériser la totalité des relations *mort(personne)* lorsque *_Tmort* est suivi d'une préposition. Ces patrons sont illustrés dans les tableaux 6.9 à 6.13 (excepté les hapax).

Patron	Fréq.	Exemple
[_Tmort _prep:de _pers]	47	mort de Jimi Hendrix
[_Tmort _prep:de _det _Ffamille]	24	mort de sa mère
[_Tmort _prep:de _locorg]	10	mort de Molière
[_Tmort _prep:d' _pers]	8	mort d'Al Mansur
[_Tmort _prep:d' 1 _action _subs]	3	mort d'un commis voyageur
[_Tmort _prep:de _det _range_objet (_Ffamille)]	3	mort de ce dernier
[_Tmort _prep:de _ens_val (_Ffamille)]	2	mort de quatre jeunes filles
[_Tmort _prep:de _det _fap]	2	mort de sa fille Charlotte Bonaparte

Tableau 6.9 – Patrons définis pour la relation mort(personne)

Patron	Fréq.	Exemple
[_Tmort _prep:en _time]	42	mort en mai 2007
[_Tmort _prep:pendant _det _subs]	3	pendant son mandat
[_Tmort _prep:en _NN]	3	mort en cours de mandat
[_Tmort _prep:dans _loc_adv (_det _time _prep _time)]	2	mort dans la nuit
[_Tmort _prep:en _pval]	2	mort en 933
[_Tmort _prep:vers _pval]	2	vers 265

Tableau 6.10 – Patrons définis pour la relation mort(date)

Patron	Fréq.	Exemple
[_Tmort _prep:à _loc]	38	mort à Berlin
[_Tmort _prep:à _locorg]	3	à Syracuse
[_Tmort _prep:à _pers]	3	mort à Florence

Tableau 6.11 – Patrons définis pour la relation mort(lieu)

Patron	Fréq.	Exemple
[_Tmort _prep:d' 1 _subs:^personnage]	10	mort d'une leucémie
[_Tmort _prep:d' 1 _NN:^personnalité]	9	mort d'une crise cardiaque
[_Tmort _prep:de _subs:tuberculose]	4	mort de tuberculose
[_Tmort _prep:pour _det _loc]	4	mort pour la France
[_Tmort _prep:de _det _subs]	3	mort de la rage
[_Tmort _prep:dans 1 _NN:accident de la route]	2	mort dans un acident de la route
[_Tmort _prep:par _subs]	2	décédé par suicide

Tableau 6.12 – Patrons définis pour la relation mort(cause)

Patron	Fréq.	Exemple
[_Tmort _prep:à _det _AAge _prep _age]	5	mort à l'âge de 25 ans
[_Tmort _prep:à _annee_dur]	4	mort à 18 ans

Tableau 6.13 – Patrons définis pour la relation mort(âge)

Un des intérêts de cette méthode est de pouvoir désambiguïser et corriger une EN en contexte. Par exemple, le patron [Tmort prep:à X] montre que si *X* est un Npr, il doit s'agir d'un lieu ; par conséquent lorsque c'est une entité-R de type Personne, c'est forcément une erreur (*mort à*

Florence). C'est une des raisons de la dispersion des patrons. Toutefois, la précision des entités-R de Rnc permet aussi de réduire cette dispersion : l'entité-R *_AAge* nous a permis d'identifier une relation que nous n'avions pas prévue, l'âge de la mort. Ces patrons ont également un potentiel pour mettre à jour de nouvelles entités-R comme on peut l'observer pour la relation *mort(cause)* : le patron [_Tmort prep:d' 1 _subs] montre que dans la majorité des cas (sauf des noms comme *personnage*), le substantif est un nom de maladie (*leucémie*, *tuberculose*) ou d'incidents (*accident de la route*, *crise cardiaque*). Tous les arguments des relations ne sont pas des Npr mais les mêmes contraintes sémantiques s'appliquent.

Pour résumer, ces cinq relations sémantiques couvrent 295 occurrences de la catégorie *_Tmort*, soit 37% des occurrences totales de la catégorie [_Tmort] et 93% des séquences [_Tmort _prep]. Toutes ces relations semblent pertinentes. Si l'on veut pouvoir les réunir dans le cadre déclenché par une occurrence de l'entité-R _Tmort, il faut explorer d'autres positions, par exemple à gauche et donc agrandir la fenêtre d'analyse. La détection de relations devient alors problématique, car plus on agrandit la fenêtre d'analyse, plus le type d'éléments rencontrés varient.

Un second problème est propre aux grammaires régulières qui imposent un ordre de détection fixe. Or, les relations que nous avons extraites occupent toutes la position R1 et elles peuvent apparaître dans un ordre variable, comme par exemple, /mort personne temps/ et /mort temps personne/. Dans un tel cadre, il semble qu'on n'ait d'autre choix que de lister les parcours possibles d'un automate, ce qui augmente le nombre de patrons ou de règles. Pour progresser et gagner en généricité, on peut s'appuyer sur un traitement syntaxique.

6.2.2. De l'apport d'une analyse en constituant pour l'EI

Deux types d'analyse syntaxique majeures peuvent être envisagées : l'analyse en dépendance que nous avons déjà pratiquée (cf. *infra* chapitre 3) et l'analyse en constituants. L'analyse en dépendance est plus complexe et suppose souvent une analyse en constituants pour déterminer les frontières des syntagmes (qu'elle soit explicite ou implicite par succession de dépendances). Nous avons donc opté pour l'ajout d'une couche d'analyse en chunk syntaxique sur les sorties du système Rnc.

L'outil que nous avons exploité est une version de Logus ([Villaneau, 2003] ; [Villaneau et al., 2007]), duquel nous avons uniquement utilisé les traitements syntaxiques. Cet analyseur, que nous nommerons LoRit, a considérablement été amélioré pour gagner en généricité et s'adapter aux sorties du système Rnc, par analyse de corpus. LoRit est un système à base de règles implémenté en C++, qui manipule les informations que nous avons exploitées dans la partie précédente pour l'extraction de relation : les formes, les entités-R et la position.

L'objectif de l'analyse en chunk est de regrouper les arbres d'entités-R en arbres plus complexes pour identifier les syntagmes nominaux (GN) et syntagmes nominaux prépositionnels (GNP), les syntagmes verbaux (GV) et les syntagmes verbaux prépositionnels (GVP) ainsi que les syntagmes adjectivaux en position attributive (GVADJ). Ces chunks sont associés à la catégorie majeure : un verbe dans le cas des groupes verbaux, un lieu dans le cas d'un GNP regroupant les entités-R [préposition déterminant lieu]. Pour illustrer le traitement effectué par LoRit nous reproduisons ici l'exemple (196) dont l'arbre est présenté dans la figure (6.7) (à comparer avec la figure (6.2) ; cf. *infra* p 164).

(196)Patricia Highsmith est morte le 4 février 1995 dans un hôpital de Locarno

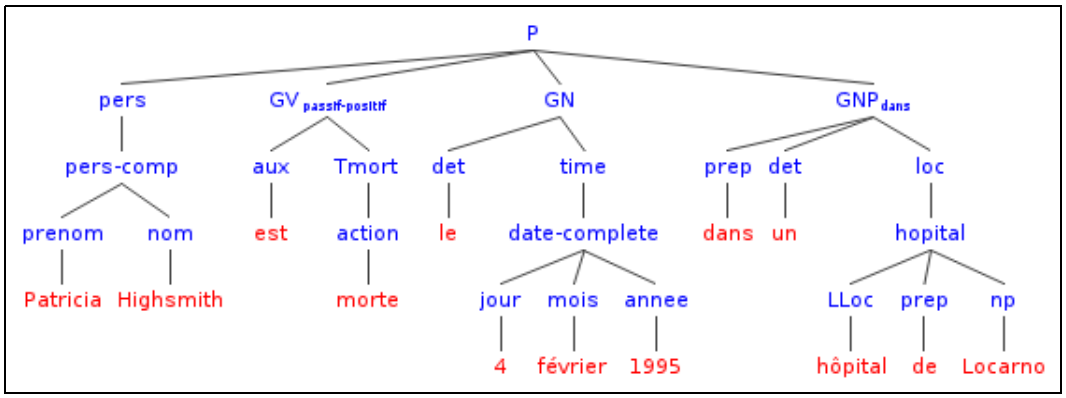


FIG 6.7 – Représentation arborescente après chunking de l'exemple 1.

Dans cet exemple, le nombre de nœuds de l'arbre de la phrase est divisé par deux sans perte d'information : la catégorie et la forme de la tête sont conservées et il est toujours possible de parcourir les fils d'un nœud pour obtenir des informations plus détaillées. Les nœuds de chunks grammaticaux possèdent en sus des attributs concernant la voix et le mode du verbe (pronominal, passif, actif) permettant de distinguer rapidement les différentes réalisations verbales. Ces attributs sont résumés dans le tableau (6.14).

Attribut	**Valeur**	**Type de Chunk**				
		GV	**GVP**	**GVADJ**	**GN**	**GNP**
Tête	*entité*	X	X	X	X	X
Voix	*actif, passif, pronominal*	X	X			
Polarité	*négatif, positif, restrictif*	X	X	X		
Modalité	*forme du modal*	X	X			
Clitique	*forme du clitique*	X	X			
Préposition	*forme de la préposition*		X			X

Tableau 6.14 – Attributs des chunks du système LoRit

Pour illustrer l'intérêt du chunking, nous avons mené une analyse du verbe *élire*, pour lequel nous espérions extraire des relations entre une personne et une fonction professionnelle. Nous avons extrait toutes les formes verbales qui répondaient à l'expression régulière /[^\w][ré]?élue?s?\W*?/ qui étaient toutes reconnus comme chunk verbal (GV) ou chunk verbal prépositionnel (GVP). Les formes obtenues sont résumées dans le tableau (6.15).

Forme verbale	Fréq.
est élu	177
est réélu	31
fut élu	18
est ensuite élu	4
fut réélu	4
a été élu	4
avoir été élu	3
avait été élu	2
est élue	2
ne sera pas réélu	2
ont élu	2
n' est pas élu	2
sera élu	2
est réélue	1
est finalement élu	1
socialiste est élu	1
était rapidement élu	1
sont désormais élus	1
est finalement largement élu	1
ne sont pas réélus	1
n' a jamais été élu	1
est néanmoins élu	1
fut toujours réélu	1
à élucider	1
est d'ailleurs élu	1
ne sera pas élu	1
d' avoir éludé	1
sont réélus	1
sera facilement élu	1
fut largement élu	1
ont été aussi élus	1
étaient auparavant élus	1
d' éluder	1
avait alors été élu	1
avait été élue	1
fut finalement élu	1
sont élus	1
fut ensuite élue	1
TOTAL	278

Tableau 6.15 – Formes extraites pour la détection de la relation de Fonction

Trois formes ne sont pas contrôlées par l'expression régulière (en rouge), et ont été supprimées. Après étude des collocats de ces formes verbales en position R1, nous avons identifié 7 catégories du cadre qu'il définit. Ces relations sont indiquées dans la figure (6.8) avec les patrons associés.

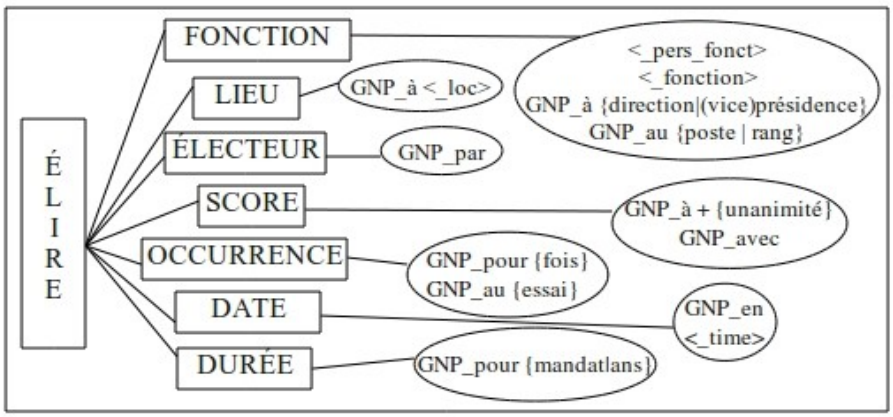


FIG 6.8 – Catégories sémantiques (encadrées) associées au verbe élire en position R1 liées à des chunks (englobés) Légende : <_Entité-R>, {Forme}

Le nombre de patrons est drastiquement réduit, comme on peut l'observer dans le tableau (6.16).

Patrons par catégorie sémantique	Fréq.
FONCTION	**81**
<_fonction>	37
<_fonc_pers>	37
GNPà {direction|(vice)présidence}	7
GNPau {rang|poste}	2
DURÉE	**40**
Gnpour {mandat|ans}	40
INSTITUTION	**26**
GNPà [reste]	20
GNPau [reste]	6
DATE	**20**
GNPen	11
GNP<_time>	9
SCORE	**4**
GNPà {unanimité}	1
GNPavec	3
OCCURRENCE	**3**
GNPpour {fois}	2
GNPau {essai}	1
LIEU	**2**
GNPà {Paris}	2
TOTAL	**176**

Tableau 6.16 – Fréquence des patrons définis, classés par catégories sémantiques associées à élire

Hormis les 23 occurrences de ponctuation en position R1, on couvre, avec ces patrons, 70% des arguments avec ces catégories sémantiques. Une plus grande difficulté se pose pour l'identification de l'élu, car la majorité des unités linguistiques en position L1 (à gauche) sont des pronoms personnels sujets (la pronominalisation ; cf. *infra* 4.1.1) ; par conséquent ils sont assujettis à une référence anaphorique, qui paraît peu utile dans un contexte d'EI, si on ne résout pas la référence. 151 collocats en position L1 sont des pronoms dont 149 pronoms sujets (dont 143 *il*, en excluant le pronom anaphorique *y*). 69 collocats sont néanmoins des personnes correctement identifiées par Rnc, ce qui fait 218 occurrences couvertes par ces deux entités-R, soit près de 80% des collocats en L1. Pour le reste il s'agit de conjonction verbale, de prépositions non rattachées au verbe, de ponctuation, etc. On peut donc concevoir un algorithme se basant sur des automates en cascade qui détectent pour une même séquence l'argument gauche et droit de façon à lier les formes du verbe *élire* à ses arguments.

Nous avons étudié la liste des verbes associés à la catégorie Personne pour identifier d'autres cadres sémantiques potentiels. Le tableau (6.17) fait figurer la fréquence d'association des 20 verbes les plus fréquemment associés à l'entité-R *_pers* dans une phrase, dans la séquence où *_pers* est en L1 et dans la séquence où il est en R1. On constate que pour la majorité de ces formes verbales, la position privilégiée est L1.

Verbe	phrase avec _pers	[_pers verbe]	[verbe + _pers]
fait	521	58	1
a	405	63	17
devient	382	70	3
voir	233	1	6
commence	188	32	0
prend	177	32	0
obtient	177	21	0
rencontre	163	21	81
meurt	145	34	2
fit	142	31	1
décide	137	44	0
avait	132	27	0
publie	129	23	4
joue	118	10	1
met	114	29	0
reste	114	22	0
reçoit	109	23	4
participe	109	14	0
passe	108	17	0
annonce	102	37	0

Tableau 6.17 – Liste des verbes fréquemment associés à l'entité-R _pers

Les seuls cas où les personnes apparaissent plus fréquemment en position R1 sont les verbes *voir* et *rencontrer*. Nous avons retenu le second pour identifier un cadre de *relations sociales* car il est plus fréquent.

Nous avons appliqué cette méthode sur quatre cadres, Mort, Naissance, Relations, Fonctions déclenché par des unités lexicales ou des expressions de ces unités lexicales (par exemple, *faire la connaissance*). Chacun de ces cadres est associé à des catégories qui sont détectées par des patrons. Les catégories correspondent à des types de question que l'on peut vouloir poser sur une personne :

- Mort
  - *Personne*, répondant à la question *qui est mort à telle date/à tel endroit ?*
  - *Date,* répondant à la question *quand est morte telle personne ?*
  - *Lieu,* répondant à la question *où est morte telle personne ?*
- Naissance
  - *Personne*, répondant à la question *qui est né à telle date/à tel endroit ?*
  - *Date*, répondant à la question *quand est née telle personne ?*
  - *Lieu,* répondant à la question *où est née telle personne ?*
- Relations
  - *Rencontrant*, répondant à la question *qui a rencontré telle personne/ à telle date ?*
  - *Rencontré*, répondant à la question *qui a rencontré telle personne/ à telle date ?*
  - *Date*, répondant à la question *quand telle personne a-t-elle rencontré telle personne ?*
- Fonctions
  - *Poste*, répondant à la question *quelle fonction a exercé telle personne ?*
  - *Personne*, répondant à la question *qui a exercé telle fonction ?*
  - *Date*, répondant à la question *quand telle personne a-t-elle exercé telle fonction ?*

Les exemples (197) à (200) illustrent des cas d'extraction complète de cadre.

(197)le 17 septembre 1863, il mourut à Paris des suites d'un cancer, après une année de souffrances physiques, courageusement supportées.
MORT *: Lieu={Paris} Date={17 septembre 1863} Personne={il}*

(198)Charles Simon Favart, né à Paris le 13 novembre 1710 et mort dans cette même ville le 12 mai 1792, est un auteur de pièces de théâtre et d' opéras-comiques français
NAISSANCE *: Lieu={Paris} Date={13 novembre 1710} Personne={Charles Simon Favart}*

(199)en 1778, Smith devient commissaire aux douanes à Édimbourg
FONCTIONS *: Date={1778} Poste={commissaire} Personne={Smith}*

(200)en 1967, elle fait la connaissance de Michel Piccoli qui deviendra son mari
RELATIONS *: Date={1967} Rencontrant={elle} Rencontré={Michel Piccoli}*

Malgré les contraintes appliquées, les extractions sont souvent partielles : toutes les catégories ne sont pas toujours présentes en contexte et quelques erreurs subsistent. Les patrons que nous avons définis sont relativement simples et peuvent être approfondis. Par exemple, pour identifier les dates, nous avons uniquement employé l'entité-R *_time*. Le cadre Fonctions fait usage d'une classe d'équivalence complexe, à cause de la variété des entités-R pouvant apparaître en position R1 :

[_fonctions, _pers_fonct, _pers_act, _subs, _np, _NN, _action, _pers, _gerondif, _org, _org_prob]

Enfin, nous avons dû définir des patrons lexico-syntaxiques peu génériques incorporant la ponctuation pour extraire des informations sur le cadre Naissance, dont un automate est présenté en figure (6.9).

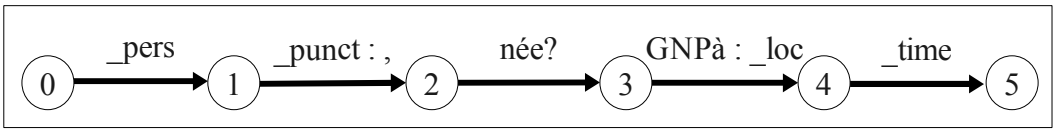


FIG 6.9 – Automate pour la détection de relations du cadre Naissance

Nous avons ainsi pu extraire automatiquement 1233 catégories pour lesquelles nous avons évalué la précision (tableau 6.18).

Cadre	Nb de catégories	Précision
NAISSANCE	37	95%
MORT	73	90%
FONCTIONS	516	75%
RELATIONS	607	70%

Tableau 6.18 – Nombre de catégories extraites et précision de ces catégories par Relation

Le problème majeur est l'évaluation du rappel, c'est-à-dire qu'on ne sait si on a bien identifié toutes les relations dans le texte. Avec une telle méthode, la construction de patrons peut s'avérer extrêmement complexe, ceci étant principalement dû aux variations discursives de surface. Nous entendons par là, les dislocations, les antépositions, les appositions, les parenthétiques, les citations, et autres phénomènes de rupture de la structure linéaire qu'une grammaire régulière peut difficilement traiter. Par exemple, il est courant dans les biographies que les dates de naissance et de mort soient incluses dans des parenthétiques qui séparent ces informations de la personne dont il est question comme en (201).

(201)Raymond Lulle (Ramon Llull en catalan) (né v. 1235 à Palma de Majorque mort en 1315) était un laïc proche des franciscains (peut-être appartint-il au Tiers Ordre des Mineurs), philosophe, alchimiste, poète, mystique et missionnaire catalan du XIIIe siècle, descendant d' une famille noble catalane.[44]

Il semble que le seul moyen, dans le formalisme des grammaires régulières, soit d'intégrer les signes de ponctuation comme états de l'automate, au même titre qu'une entité-R, rendant les règles très lourdes à créer et peu génériques. S'agit-il d'une limite de ce formalisme de patrons lexico-syntaxiques ?

En conclusion, la construction de grammaires pour l'extraction de relations sémantiques dans des cadres fait face à trois problèmes majeurs :

- Une même catégorie sémantique s'exprime par diverses entités-R ou structures : nous y avons répondu en partie en utilisant le chunking.
- L'ordre dans lequel les éléments sont présentés en discours complexifie le nombre de patrons de règle : nous avons proposé un traitement module par module en cascades (un module par catégorie) qui peut paraître insuffisant.
- La détection de relations pose des difficultés de taille à cause de la variation des phénomènes que l'on peut rencontrer entre le noyau du cadre et son argument : comment gérer les relations à longue distance ?

44 Cet exemple a été reformaté et normalisé dans la version actuelle de l'encyclopédie, dans laquelle ces problèmes ne se posent plus (modification du 13 Octobre 2006).

6.3. Bilan

Ce volet nous a fourni l'occasion de décrire en détail le domaine d'application privilégié de notre travail sur l'analyse de relations sémantiques, l'Extraction d'Information. Notre présentation s'est focalisée sur la notion d'Entités Nommées (EN). Nous avons montré que les campagnes initiales masquaient un certain nombre de phénomènes récemment mis en relief et qui constituent des limites des systèmes actuels. On constate en effet des dégradations de performances des systèmes lorsque les catégories sont plus finement décrites, ou lorsqu'elles sont sujettes au phénomène de métonymie. Bien que la Reconnaissance et la Classification des EN (RCEN) ait été conçue au départ comme une technologie portable et générique, les recherches ont montré que la tâche s'avérait plus complexe lorsqu'on variait des paramètres textuels, en changeant par exemple ne serait-ce que la date du corpus [Mota & Grishman, 2008]. D'autres facteurs comme le style, le domaine, le genre, le canal ou la capitalisation sont reconnus comme des sources d'influence des performances des systèmes et posent la question de l'adaptation au cœur de la recherche.

Adapter un système ou des ressources suppose de définir un contexte de départ à partir duquel une projection peut être envisagée. Nous avons proposé de replacer l'étude des EN dans le contexte général dans lequel elles ont été définies, l'extraction de cadres d'information (templates), correspondant à des formulaires qui les organisent par rapport à un concept-pôle (événement ou scénario). En faisant écho à nos analyses sur les liens entre texte et référence, nous avons défendu l'idée que les cadres constituent des contextes discriminants pour structurer la connaissance (catégories et relations sémantiques) autour d'un concept-pôle. Ce sont aussi des outils sémantiques pertinents pour les Systèmes de Question Réponse, qui constitue notre cadre de recherche : l'analyse d'une question factoïde peut alors être appréhendée comme l'identification d'une relation sémantique associée à une EN [Srihari & Li, 2000] dans un cadre d'information déclenché par une unité lexicale. L'apport de ressources sémantiques comme FrameNet[45] pour un SQR n'est pas encore tranché notamment parce que leur intégration dans un SQR est une tâche complexe (voir [Narayanan & Harabagiu, 2004] et [Ofoghi, 2009]). De telles ressources n'existent pas encore pour le français (mais voir [Pado & Pitel, 2007]), et nous avons déjà montré quelques intérêts et difficultés d'application des cadres de FrameNet en corpus (cf. *infra*, 4.3.1).

Étant donné notre approche linguistique, notre intérêt va vers les systèmes symboliques qui ont l'avantage d'être transparents, à l'inverse des systèmes supervisés. Néanmoins, notre travail sur la modélisation linguistique du contexte se situe en amont de l'implémentation et nos résultats peuvent être profitables pour toutes les familles de systèmes.

Pour expérimenter l'extraction de relation sémantique sur des systèmes symboliques, nous avons proposé une étude de faisabilité sur l'extraction de cadres biographiques. De nombreux systèmes ont été proposés pour extraire des informations biographiques. Notre approche se rapproche des travaux de M. Geierhos et ses collègues [Geierhos et al., 2009] dans l'utilisation de patrons lexico-syntaxiques pour acquérir semi-automatiquement des relations sémantiques. Notre objectif était d'une part, d'appliquer une méthode ascendante d'analyse de corpus pour faire émerger ces relations et d'autre part, de caractériser les difficultés rencontrées par les systèmes symboliques. Ceci nous a permis d'identifier la contrainte d'ordre et la distance entre deux éléments liés par une relation comme principaux problèmes. Si le premier peut être en partie solutionné par une modularisation des règles (un module dédié à chaque relation sémantique et invoqué pour chaque position), le second est plus complexe à résoudre à cause de la variété des éléments (du point de vue de la taille ou de la nature linguistique) qui s'insèrent en position intermédiaire.

45 voir également PropBank et Verbnet ; cf. : http://verbs.colorado.edu/~mpalmer/projects.html

QUATRIÈME VOLET

SEGMENTATION DISCURSIVE DE SURFACE

Les différentes limites des grammaires régulières pour la détection de relations sémantiques que nous avons exposées au chapitre précédent nous ont conduit à aborder la question de l'extraction de relations sémantiques sur le plan discursif. Dans ce volet, nous proposons un modèle de segmentation qui permet de gérer certains phénomènes de variation discursive, tout en conservant le formalisme des grammaires régulières.

L'objectif d'une segmentation discursive de surface est simple : elle consiste à définir des unités plus larges que le mot à partir d'indices, comme les virgules. Cette définition d'une apparente simplicité masque la complexité des phénomènes que ces unités recouvrent : qu'y trouve-t-on ? Correspondent-elles à des unités sémantiquement cohérentes ? Comment interagissent-elles ? Quel(s) intérêt(s) présentent-elles pour l'extraction d'information ?

Le chapitre 7 est consacré à l'étude de ces structures ainsi qu'à la description du segmenteur qui les détecte automatiquement.

Dans le chapitre 8, nous montrerons comment ce système peut répondre à la problématique de la détection de relations à longue distance.

7. Segmentation de surface

Pour introduire la problématique de la segmentation discursive et présenter le modèle que nous avons défini, nous traiterons dans un premier temps d'un type d'extraction d'information particulier, l'extraction de citations. La segmentation discursive de surface sera abordée en seconde partie. Elle nous permettra de définir les exigences auxquelles sera soumis le segmenteur décrit en troisième partie. Le chapitre s'achève sur une description des segments que permet d'obtenir ce système.

7.1. L'extraction de citation

L'observation de la liste ordonnée des verbes d'un corpus peut être un premier indicateur de la nature des informations que l'on va trouver fortement représentées. Comme nous disposons de deux corpus dans le même format, nous pouvons comparer les formes verbales les plus courantes. Les tableaux (7.1) et (7.2) font figurer les 30 formes verbales les plus fréquentes dans les deux corpus.

Forme verbale	Fréq.
fait	650
a	519
devient	452
voir	279
commence	220
prend	217
obtient	191
est élu	174
rencontre	171
meurt	163
fit	161
permet	159
reste	156
ont	156
décide	153
faut	148
avait	143
met	143
publie	142
joue	133
sort	130
compte	123
passe	122
annonce	122
reçoit	120
donne	117
a été	116
va	115
participe	112
apparaît	112

Forme verbale	Fréq.
a	27518
fait	10761
faut	8340
ont	8278
n' a pas	6963
reste	5331
avait	5231
explique	5060
compte	5005
faire	4190
vient	4037
été	4014
a été	3752
s' agit	3683
va	3608
estime	3509
semble	3388
a fait	3327
affirme	3148
être	3021
a annoncé	2954
a déclaré	2935
n' ont pas	2878
font	2782
permet	2767
sait	2470
pourrait	2436
souligne	2274
prend	2165

Tableaux 7.1-7.2 – Fréquence de formes verbales. biographies (gauche) et presse (droite)

Les fréquences ne sont évidemment pas comparables, mais on peut remarquer qu'un certain nombre de formes sont communes (en rouge, il y en a 11), et qu'elles consistent en des unités peu informatives (auxiliaires par exemple). On peut *a contrario* faire émerger des formes spécifiquement fréquentes dans chaque corpus. On observe par exemple qu'aucun verbe n'est au pluriel dans le corpus de biographies (*ont*, *n'ont pas*, *font*), que les formes sont généralement simples et qu'on trouve même des passés simples (*fit*). Certains verbes ont été employés dans les cadres précédemment présentés, mais on peut y ajouter la publication d'œuvre (*publie*, *sort*), la réception de prix (*obtient*, *reçoit*) ou encore la participation à des événements (*participe*, *joue*).
Pour le corpus de presse, on observe un nombre plus important de verbes à usage auxiliaire (*a*, *ont*, *n'a pas*, *avait*, *été*, *a été*, *être*, *n'ont pas*, *vient*, *va*), de verbes dits supports (*fait*, *reste*, *faire*, *a fait*, *font*, *prend*) et de verbes à usage épistémique ou déontique (*faut*, *compte*, *pourrait*, *semble*). Ces formes verbales peuvent prendre des sens spécifiques en contexte (interprétation causale pour le verbe *faire*, comme dans *faire faire*, interprétation de mouvement pour *vient* et *va*) et il est

également possible que certaines occurrences soient des parties de chunks verbaux incomplets voire qu'ils appartiennent à un paradigme autre que verbal, comme *été* qui peut désigner la saison estivale. Les autres verbes sont des verbes à usage citationnel : *explique*, *estime*, *affirme*, *a annoncé*, *a déclaré*, *souligne*, dont une forme uniquement figure dans le corpus de biographie (*annonce*). Ces verbes peuvent également être employés dans d'autres sens, ce qui nécessite une validation.
La forte présence de verbe citationnel s'explique par le rôle médiatif du journalisme : les articles font régulièrement intervenir des personnalités dont ils citent les paroles. Une application d'extraction d'information consisterait donc à extraire les paroles de personnes en focalisant sur ce groupe de verbes. Nous avons déjà indiqué (en 3.3.3) que les analyseurs syntaxiques n'exploitent pas les relations entre le verbe et le discours direct, ce qui modifie la distribution de fréquence des structures argumentales auxquelles ces verbes sont associés.
La forme qui entre le plus souvent en collocation avec ces verbes est le signe de ponctuation *guillemet* comme dans l'exemple (202).

(202)"Rochefort se dope aux projets surdimensionnés", sourit Emmanuel de Fontainieu

On constate qu'elle est en effet très proche du verbe en termes de distance, si on la considère comme un élément de l'empan (les signes de ponctuation sont généralement supprimés dans les analyses collocationnelles). Cet exemple instancie une structure récurrente dans les articles de presse :

[" CITATION " , Verbe Locuteur]

Syntaxiquement, il s'agit d'une construction verbale en sujet inversé, apposée à un discours direct. Sémantiquement, le rôle du sujet est la source de la citation, le Locuteur, qui peut être une personne, une organisation ou encore un support médiatique (journaux, chaîne de télévision, etc.). Nous avons cherché à extraire tous les verbes qui entraient dans cette construction en employant comme contrainte dans nos patrons tous les autres indices. On peut difficilement détecter une citation dans une grammaire régulière, puisqu'on ne peut définir par avance les éléments qui séparent le guillemet initial du guillemet final. Nous avons donc uniquement employé le guillemet final comme indice.

Les verbes obtenus ne sont pas tous typiquement des verbes de parole, mais des verbes qui vont tantôt décrire le mode de communication (*écrit*), caractériser la nature de la citation (*se souvient*, *analyse*, *regrette*), et certains usages souvent considérés comme des métaphores (*glisse*, *martèle*, *s'enflamme*). Cette simple grammaire permet d'extraire près de 30 000 occurrences de type citationnel, parmi lesquels 281 formes verbales sont complètement couvertes (ce qui représente 786 occurrences) : ces dernières sont uniquement employées dans cette structure. Les 10 formes verbales qui apparaissent le plus fréquemment dans ces patrons (Fréquence dans la structure, Fréquence globale, Proportion) sont listées dans le tableau (7.3), triées en fonction de leur fréquence d'usage dans cette structure.

verbe	Freq. Cit.	Freq. Tot.	Prop. Cit.
explique	2807	5062	0,55
a déclaré	1263	2936	0,43
affirme	1067	3148	0,34
estime	1005	3558	0,28
souligne	937	2276	0,41
dit	914	3983	0,23
précise	639	1679	0,38
assure	629	1856	0,34
a-t-il dit	605	620	0,98
a-t-il ajouté	595	603	0,99

Tableau 7.3 – Verbes entrant dans les structures citationnelles

On observe que ce type de structure est prépondérant pour certaines formes verbales (55% de la forme *explique*, par exemple). On remarque également que certaines formes (*a-t-il dit*, *a-t-il ajouté*) intègrent les pronoms qui apparaissent entre l'auxiliaire et le participe : ils sont préalablement identifiés par la passe de chunking. Ces formes sont systématiquement employées dans des structures citationnelles, l'inversion pronominale constitue donc un indice fort de désambiguïsation.

Cette caractéristique du discours de presse écrite invite alors à recomposer le cadre sémantique de ces citations en structurant plus finement ses arguments. Les patrons sont incomplets car il faut identifier le Locuteur (par co-référence ou par restriction syntaxico-sémantique), délimiter la citation et extraire éventuellement d'autres rôles, parmi lesquels, la Date, le Lieu, le Destinataire et le Support Médiatique (*sur Europe 1*). Ces rôles ne correspondent pas à une même entité-R et nécessitent la conception de lexiques appropriés. Étant donné que ces rôles peuvent apparaître dans un ordre variable après le verbe, le système d'extraction doit également relâcher ses contraintes de position, tout en poursuivant son analyse. Mais la détection de la citation est le problème majeur, car :

- la forme verbale peut se situer en incise, c'est-à-dire dans la citation.
- une citation peut s'étendre au-delà d'une phrase
- une citation peut ne comporter que quelques mots qui prennent leur sens dans la proposition qui l'intègre (dans un patron comme [X considère la Y pas toujours "Z"])
- les groupes entre guillemets ne correspondent pas toujours à des citations (mots utilisés avec précaution par l'auteur par exemple)

La citation relève de la segmentation discursive, que nous allons à présent aborder.

7.2. La segmentation discursive de surface

7.2.1. La ponctuation : Rupture ou Structure ?

La question des phénomènes discursifs de surface se révèle avec une particulière acuité dans les corpus écrits, car ces derniers regorgent de phrases complexes et d'un usage de la ponctuation sophistiqué. Pourtant, le traitement de la ponctuation en linguistique comme en TAL est très limité. Bayraktar et ses collègues imputent ce manque à l'absence de cadre théorique :

> « Until recently, punctuation has been neglected by most researchers in theoretical and computational linguistics. This is due to the absence of a concise, formal background for the abstract problem. However, once we remember that punctuation is an orthographical component of written language, we see that research on punctuation makes reasonable sense. Accordingly, interest in the subject rose within the last decade because it has been realized that a fuller understanding and processing of written language is quite impossible without taking punctuation into account. Although punctuation was originally invented as a device for reflecting intonation in written text, it is now a linguistic "system on its own right" (Nunberg 1990: 9) » [Bayraktar et al., 1998 : 1]

L'étude de la ponctuation relève bien de la linguistique, ce qui soulève un certain nombre de questions. Peut-on parler de polysémie des signes de ponctuation ? Comment les classer ? Peut-on identifier des règles qui leur sont propres ? Ces règles relèvent-elle d'un style particulier ? Peut-on identifier des régularités ? Que nous enseignerait un traitement quantitatif de ces signes et que compter ?

Si elle peut servir à transcrire des propriétés de l'oral (intonation des points d'exclamation par exemple), la ponctuation est essentiellement considérée comme un indice de rupture, à l'image du point qui peut délimiter des phrases. Comme l'indique G. Purnelle à propos de la classification de Tournier, c'est leur fonction principale :

> « Tout d'abord, comme le montre la classification de Claude Tournier, tous les signes de ponctuation assument, apparemment, au moins une fonction de séparation (de délimitation) ; c'est donc leur fonction principale : un point sépare deux phrases, une virgule deux membres de phrase ou deux syntagmes ; le deux-points sépare phrase citante et phrase citée. Quant au 3e groupe, on notera que marquer une interruption ou une inclusion, comme le font les signes doubles, c'est également délimiter, séparer le texte inclus de ce qui l'entoure. » [Purnelle, 1998 : 214]

Les signes doubles s'opposent aux signes simples par le fait qu'ils se répondent et délimitent une unité au sein d'une phrase. Pour illustrer ce phénomène de rupture, prenons l'exemple d'une phrase d'un article du corpus LeMonde, dans laquelle la relation entre un verbe et son sujet est interrompue (203).

(203)Le procès de Mijailo Mijailovic, meurtrier présumé de l'ex-ministre suédoise des affaires étrangères, Anna Lindh, s'est ouvert, mercredi 14 janvier à Stockholm, quatre mois après le drame.

Cette phrase comporte 5 virgules, dont 3 qui précédent le verbe principal (*s'est ouvert* ; c'est d'ailleurs le seul verbe conjugué). On observe que le verbe est lui-même isolé entre deux virgules,

qui délimitent deux appositions. Ce sont des signes doubles, mais on doit remarquer que la virgule initiale de la séquence *Anna Lindh* est aussi la virgule finale de l'apposition qui précède.
Pour identifier la relation qui lie *procès* à son verbe, il faut pouvoir gérer en amont les deux appositions qui précèdent le verbe.

- Une stratégie palliative peut consister à supprimer les virgules et repérer le premier SN à gauche : ceci conduira à un mauvais choix de sujet (*Anna Lindh*).
- D'après nos observations, une apposition se rattache généralement à l'élément qui le précède ou encore à la proposition globale (dans le cas d'une apposition temporelle par exemple). Dans ce cas, il s'agirait de rattacher *Anna Lindh* au segment qui le précède, en faisant de même pour *meurtrier présumé de l'ex-ministre suédoise des affaires étrangères*. Cette opération ne serait pas suffisante, car l'apposition située entre *Le procès de Mijailo Mijailovic* et *Anna Lindh* les met en relation.
- Il faut donc rattacher *Anna Lindh* à un élément du segment qui le précède, plus exactement à *l'ex-ministre suédoise des affaires étrangères*. Les segments ainsi enchâssés, on peut alors récupérer la tête du syntagme à gauche qui est *procès* comme candidat sujet.
- Une séquence de mots entre virgules n'est pas nécessairement une apposition, puisque si c'était le cas, le verbe lui-même serait en apposition ; or, c'est le verbe principal.

Comme nous l'avons déjà noté, l'intégration des virgules dans une grammaire alourdit le traitement. Une des solutions serait de lister les structures alternatives pouvant apparaître entre un verbe et son argument. En les considérant comme optionnelles, on allège partiellement les règles, comme pourrait le permettre une expression régulière illustrée en (204) et (205).

(204)GN (virgule Adverbe virgule) ? Verbe
(205)GN (virgule Nom virgule Npr virgule) ? Verbe

Une telle expression définit néanmoins a priori les éléments pouvant apparaître entre virgules : quel mode de représentation choisir, syntaxique (*adverbe*, *Npr*) ou sémantique (*temps*, *Personne*) ? Or on ne connaît pas exactement les formes que peuvent prendre ces appositions. Par ailleurs, cette solution doit lister les alternatives : elle n'allège que partiellement les règles.

La majorité des analyseurs symboliques syntaxiques (Syntex par exemple), multiplient le nombre de règles pour détecter des sujets à longue distance et sont ainsi limités par le nombre d'enchâssement successifs qu'ils peuvent gérer. Cette phase n'est en général pas décrite explicitement.

La ponctuation n'est pas uniquement un indice de rupture syntaxique, elle joue un rôle de structuration dans la présentation de l'information. Nous avons déjà indiqué l'intérêt qu'il y a à détecter des citations entre guillemets pour l'extraction d'information, mais les structures appositives (206) et parenthétiques (207) relèvent également de relations sémantiques.

(206)le secrétaire d' Etat aux PME, Renaud Dutreil , s' était ainsi vu convoquer par le directeur du cabinet du premier ministre, Michel Boyon.
(207)Eric Tanguy (né en 1968) passait alors pour un nouveau Dusapin

En (206), les appositions renferment les Npr des référents désignés par le SN qui les précède. Cette structure peut être inversée : la fonction professionnelle apparaîtrait en apposition. Ces deux structures semblent similaires mais elles diffèrent par la forme que prend la ponctuation finale :

virgule pour *Renaud Dutreil*, point pour *Michel Boyon*. L'identité du signe de ponctuation n'est donc pas un critère suffisant pour détecter la totalité de ces structures.
En (207), on observe que c'est un cadre de Naissance qui est inclus dans une parenthétique : il contient le nœud déclencheur (*né*) avec la date et est séparé de l'argument Personne.
L'identification de telles structures peut donc enrichir la détection de relations sémantiques.

Il semble qu'il soit nécessaire de séparer le traitement discursif du traitement syntaxico-sémantique. En TAL, la question des relations sémantiques à ce niveau discursif est rarement envisagée du point de vue de l'extraction d'information. Les systèmes de TAL qui prennent en compte la ponctuation, en révélant ainsi son utilité, sont

(i) orientés vers la détection de relations de discours [Marcu, 2000]

(ii) orientés vers la détection de quelques patrons lexico-syntaxiques, notamment pour la détection de structures énumératives ([Marti A. Hearst, 1992], [Morin, 1998]).

Pour traiter ces occurrences, nous avons exploré la voie suivante : découper la phrase en segments, les catégoriser et les mettre en relation.

7.2.2. Type de segment et classe de frontière

Pour concevoir un système qui puisse gérer correctement les variations discursives, nous devons tout d'abord définir ce qu'est un segment. Nous proposons trois types de segments : un segment se caractérise par sa frontière gauche et sa frontière droite et peut être équilibré ou déséquilibré.

1. Dans un corpus de presse, le segment le plus long est l'article (pouvant parfois atteindre des dizaines de milliers de mots). Les frontières fondamentales sont donc « Début de texte » et « Fin de texte ». Ce sont, du point de vue de la détection des relations sémantiques qui nous intéressent, ce que nous allons qualifier de frontière forte : aucune information n'est à chercher au-delà de cette frontière, les relations doivent unir les éléments situés entre des frontières fortes. La classe des frontières fortes que nous avons définie est composée des signes suivants :

   ! . ? ; : ...

   Le point est le signe le plus ambigu, il est régulièrement employé comme signe d'abréviation (*S. T. Coleridge*), ce qui justifie l'usage d'un système de RCEN. La frontière initiale correspond à la frontière finale du segment précédent. Ces frontières séparent des segments équilibrés qui sont proches des phrases (des erreurs peuvent exister, le statut de phrases de certains segments averbaux ou titres est discutable, etc.). Le segmenteur doit donc parvenir à représenter la phrase de l'exemple (208) par un arbre illustré en figure (7.1).

   (208)M. Deleuze est en Croatie.

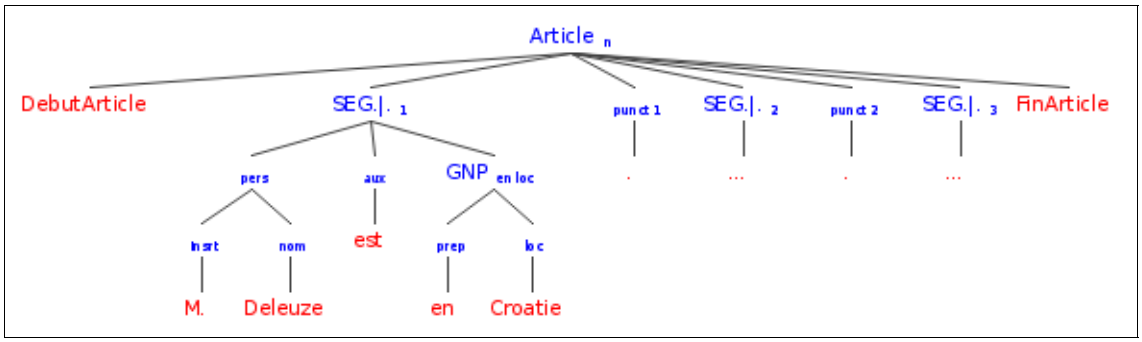


FIG 7.1 – Arbre syntaxique de l'exemple (208)

Dans cette figure, les segments sont des nœuds (nœud *SEG* accompagné des frontières .|.) qui contiennent les entités-R et chunks détectés par les systèmes précédents. Les frontières n'en font pas partie : ces derniers ne jouent pas de rôle au sein du segment.

2. Les frontières semi-faibles sont des frontières englobantes que l'on peut vouloir transgresser. La classe est composée des signes suivants :

( ) " [ ]

Les parenthétiques et citations peuvent inclure des éléments très divers dont les points et les virgules. On ne peut les traiter au même niveau que les frontières fortes, car on créerait alors des segments déséquilibrés, dont les frontières seraient une parenthèse et un point par exemple. Le segmenteur devra donc identifier deux frontières semi-faibles pour créer un tel segment. La représentation attendue de la phrase de l'exemple (209) est illustrée en (7.2).

(209)Les économistes tablent sur une croissance de 4 pour cent du produit intérieur brut (PIB) aux États-Unis.

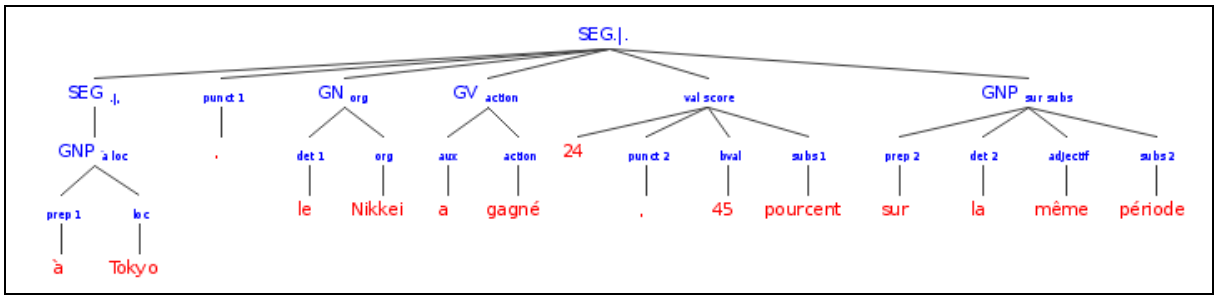


FIG 7.2 – Arbre syntaxique de l'exemple (209)

On constate qu'ici, la parenthétique fait partie d'une phrase, qui est déjà un segment.

3. La virgule est le prototype même de frontière faible, elle sépare des éléments sémantiquement associés comme la relation sujet. Elle est aussi sujette à ambiguïté, comme lorsqu'elle sépare un nombre d'une décimale. Elle peut aussi isoler deux propositions qui n'entretiennent que des relations discursives, comme en (210).

(210)La pluie commence à tomber, les gens courent s'abriter.

Dans ce cas la virgule partage le même rôle que les conjonctions de subordination (*parce que*, etc.). Comme nous l'avons vu, la virgule peut aussi isoler des groupes en apposition, auquel cas elle remplit plutôt la fonction d'une frontière semi-faible. Il n'est pas possible de savoir par avance la fonction qu'entretiendront des segments séparés par des virgules : elles peuvent séparer des verbes en énumération, des syntagmes, comme des propositions. Les frontières faibles regroupent donc des formes aux usages très disparates qui demandent un traitement ultérieur. Par conséquent nous y avons inclus tous les signes de ponctuation et formes qui permettent d'isoler des propositions ou des groupes plus « faibles » comme les syntagmes au sein d'une phrase. Seules les frontières en filiation directe avec la racine d'un segment fort dans l'arborescence de l'article donnent lieu à une segmentation en frontière faible. Nous ne considérons pas comme frontière faible le tiret, qui tantôt joue le rôle de trait d'union (dans de nombreux mots composés), tantôt de séparateur, équivalent à la parenthèse. Le problème est que le tiret peut initier un segment qui ne s'achève que par un point. La classe des frontières faibles contient donc les virgules, complétée par les conjonctions de subordinations et connecteurs :

VIRGULE, que, si, puisque, quand, lorsque, quoique, parce que, dont, comment, pourquoi, alors que, ainsi que, ce qui, dès que, comme, où, qui, car, puis, mais, alors, etc.

Ce choix est discutable du point de vue de la formalisation, mais un traitement convenable nécessite une analyse contextuelle complexe, ce qui ne peut être révélé que par une étude détaillée de chaque membre de cette classe. Par ailleurs, la segmentation n'est pas une opération irréversible : rien n'empêche de l'affiner ultérieurement.

Les frontières faibles peuvent créer des segments déséquilibrés : on peut trouver des segments à initiale forte mais finale faible ou l'inverse. Le segmenteur doit pouvoir représenter la phrase de l'exemple (211) dans la figure (7.3).

(211)à Tokyo, le Nikkei a gagné 24,45 pour cent sur la même période.

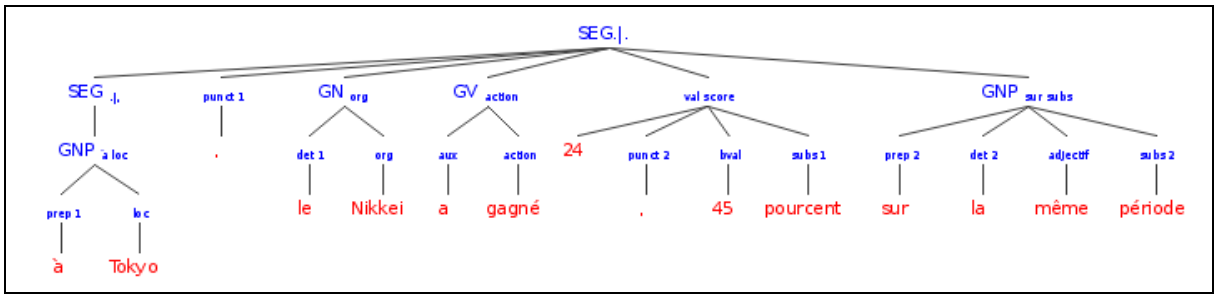


FIG 7.3 – Arbre syntaxique de l'exemple (211)

On observe que la première virgule a servi d'indice de segmentation à l'inverse de la seconde, qui fait partie de l'entité-R *_val_score* et n'est donc pas fille immédiate de la racine. On trouve là un autre avantage de l'usage d'un système de RCEN.

La définition des segments que nous proposons est discutable à bien des égards, mais elle a l'avantage d'être opérationnelle et d'imposer un minimum d'a priori sur la nature des frontières et des segments qu'elles séparent. Pour y parvenir, il faut disposer d'un segmenteur qui implémente cette formalisation.

7.2.3. Le segmenteur

Le principe du segmenteur est de créer des sous arbres dans l'arbre du document. Il prend du texte en entrée, mais a été optimisé pour un format XML tel qu'obtenu en sortie des systèmes Rnc et LoRit. Ces systèmes permettent, comme nous l'avons évoqué pour les abréviations et les nombres à décimales, de limiter l'ambiguïté de certains signes de ponctuation, ce qui améliore ses performances et évite un traitement trop lourd. L'algorithme principal consiste donc à parcourir l'arbre du document en largeur d'abord sans explorer les chunks ni les entités-R, sauf s'il s'agit de frontières (comme dans le cas des conjonctions). Un second algorithme permet de traiter les cas d'imbrication de segments (segment de type parenthétique dans une phrase par exemple). Comme l'imbrication peut être complexe, le traitement est récursif.
Le segmenteur opère en trois phases :

1. Identification des segments semi-faibles
2. Identification des segments forts
3. Identification des segments faibles

1. Les segments semi-faibles désignent les parenthétiques, les citations et les séquences entre crochets qui nécessitent d'être équilibrées. Cette segmentation est prioritaire parce que ces segments peuvent contenir les deux autres types de segments. Elle opère sur deux niveaux d'imbrication.

   * L'algorithme doit d'abord identifier une des frontières initiales comme *(*, *"* ou *[*, c'est-à-dire un nœud de l'arbre qui est une entité-R de type *_punct* (ponctuation) dont la feuille correspond strictement à l'une de ces frontières. Il recherche ensuite la frontière fermante correspondante *)*, *"* ou *]*.

   * Dans le cas où il rencontre une autre frontière semi-faible différente de celle qu'il a déjà trouvée, cette dernière devient prioritaire pour la recherche de fermeture. L'identification de la frontière finale de la première n'intervient que lorsque celle de la seconde est identifiée, ou lorsqu'une nouvelle frontière est identifiée (ce qui signifie que cette dernière opération a échoué).

   Nous ne permettons pas que deux segments semi-faibles de même nature soient imbriqués, comme une citation dans une autre citation. Une telle omission occasionne nécessairement des erreurs, mais un traitement satisfaisant nécessiterait de prendre en compte la totalité des frontières semi-faibles d'un document et d'effectuer des choix qui seraient moins génériques et sensibles, étant donné que certains signes peuvent disparaître dans les pré-traitements ou tout simplement avoir été omis par l'auteur de l'article.

   Les segments identifiés sont rattachés au nœud racine de l'article.

2. Les segments forts sont identifiés sous ce nœud et dans les segments semi-faibles lorsqu'ils sont rencontrés. Le segmenteur crée un segment dès qu'il repère un des signes de ponctuation forts. Les frontières initiales et finales de l'article sont ajoutées pour le cas où

les articles ne s'achèvent pas par un de ces signes.

3. L'identification des segments faible est effectuée de la même manière que pour les précédents, récursivement sous la racine ainsi que tous les segments rencontrées. Les frontières faibles sont stockées dans un lexique et la règle de détection ignore la casse.

À chaque fois qu'une frontière est identifiée, les nœuds stockés au cours du parcours (excepté les frontières) sont détachés de la racine et ajoutés comme filles du nouveau nœud de type Segment : ce dernier est rattaché à droite de la frontière initiale ou à gauche de la frontière initiale s'il s'agit d'un début d'article, ou encore à gauche de la frontière finale s'il s'agit de la fin de l'article. L'application d'une telle segmentation de surface sur un article de presse résulte en une représentation extrêmement fragmentée qui demande à être analysée. Dans le chapitre suivant, nous proposerons des exemples de règles de Segment.

Lorsque la phase de segmentation est achevée, l'arbre du document est imprimé dans un fichier pour garder une image de l'arbre, avant l'application de règles de Segment. La figure (7.4) montre le résultat de segmentation de la phrase précédente.

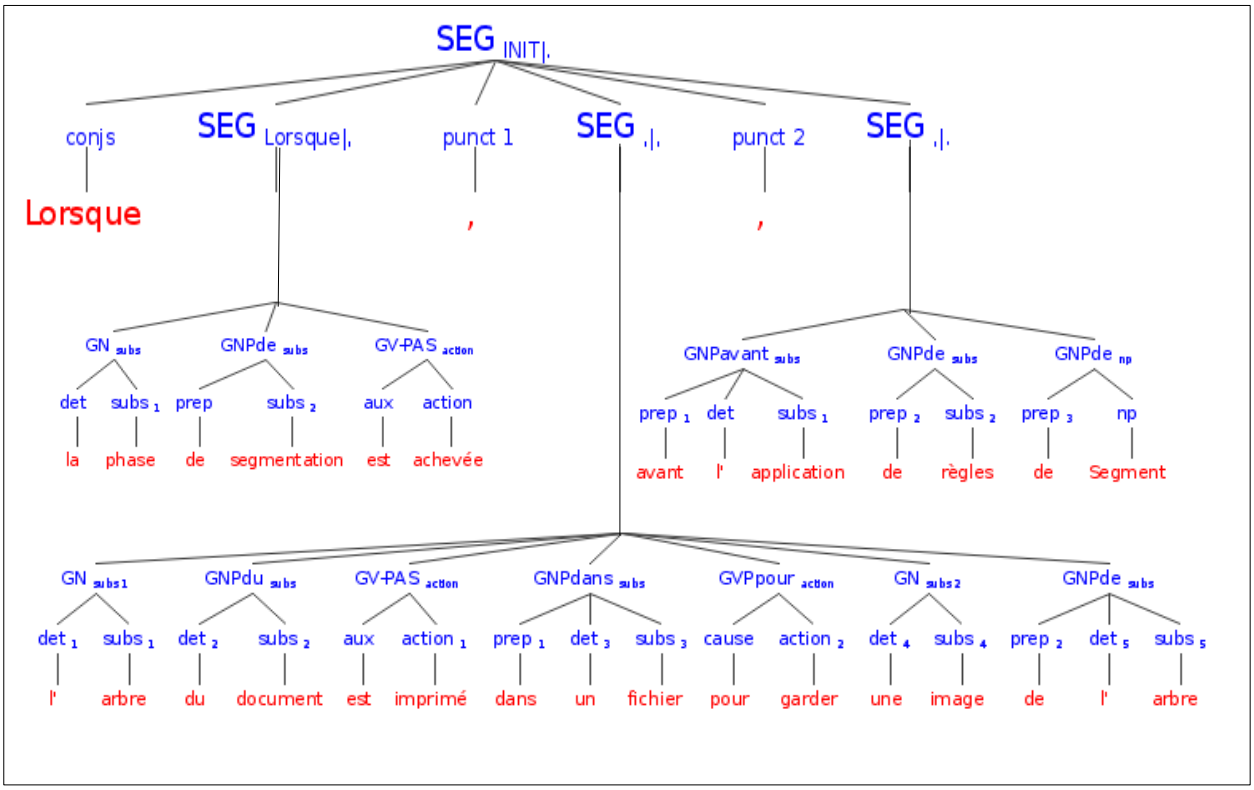


FIG 7.4 – Exemple d'arbre fourni en sortie du segmenteur

Visuellement, les segments se succèdent et sont séparés par des frontières. Lorsque plusieurs frontières se succèdent, aucun segment n'est créé. Sur la figure, les virgules permettent d'isoler deux segments qui sont des propositions, l'une subordonnée, l'autre principale. Ils peuvent donc être analysés indépendamment. Par contre, le dernier segment contient un complément circonstanciel de temps qui ne peut s'interpréter seul. Pour chaque segment, les informations concernant ses frontières, sa taille et les nœuds sont conservées. Elles sont utilisées pour les caractériser.

7.3. Segments : une vue d'ensemble

Les signes de ponctuation sont aussi fréquents que les formes de mot (tableau 7.4).

Formes	Fréq.	Prop.
,	1533241	6,81%
de	982808	4,37%
.	909902	4,04%
la	508511	2,26%
"	408633	1,82%
le	365872	1,63%
à	358376	1,59%
et	329005	1,46%
les	306591	1,36%
des	301536	1,34%
en	240818	1,07%
du	225005	1,00%
un	163862	0,73%
a	157251	0,70%
qui	145753	0,65%
que	144193	0,64%
dans	136519	0,61%
)	136202	0,61%
(	136040	0,60%
une	135492	0,60%
pour	129899	0,58%

Tableau 7.4 – Formes les plus fréquentes du corpus de presse

La virgule est le signe de ponctuation le plus fréquent, suivi par le point, les guillemets et les parenthèses qui en tout réalisent plus de 14% des signes du corpus. Parmi les phrases s'achevant par un point, uniquement 16% ne contiennent pas d'autre signe de ponctuation : l'adéquation [1 phrase simple = 1 proposition] est faiblement représentée, et la majorité des phrases sont complexes.

Sur le corpus de développement de 32750 articles (soit 17 millions de formes), notre module de segmentation obtient 2 413398 segments, soit en moyenne 70-80 segments par article. La taille des segments est un premier indice utile à leur caractérisation : elle est calculée en fonction du nombre de nœuds filles d'un segment (entités-R et chunks syntaxiques) que nous appellerons *éléments*. Par exemple, le nombre de segments comportant un élément est de 616284, soit environ 25% de segments créés. Le tableau (7.5) résume les caractéristiques des segments en fonction de la taille. Il indique d'une part, le nombre de segments (*NB S.*) et d'autre part le nombre d'éléments concernés (*NB E.*), ainsi que leurs proportions (*Prop.*) et proportions cumulées (*Prop. Cumul.*).

Taille	Nb S.	Nb S. cumul.	Prop. S.	Prop. S. cumul.	Nb E.	Nb E. cumul.	Prop. E.	Prop. E. cumul.
1	616284	616284	25,54%	25,54%	616284	616284	8,18%	8,18%
2	571014	1187298	23,66%	49,20%	1142028	1758312	15,16%	23,33%
3	453120	1640418	18,78%	67,97%	1359360	3117672	18,04%	41,37%
4	282114	1922532	11,69%	79,66%	1128456	4246128	14,98%	56,35%
5	181529	2104061	7,52%	87,18%	907645	5153773	12,05%	68,39%
6	114299	2218360	4,74%	91,92%	685794	5839567	9,10%	77,50%
7	72901	2291261	3,02%	94,94%	510307	6349874	6,77%	84,27%
8	45601	2336862	1,89%	96,83%	364808	6714682	4,84%	89,11%
9	28833	2365695	1,19%	98,02%	259497	6974179	3,44%	92,55%
10	17778	2383473	0,74%	98,76%	177780	7151959	2,36%	94,91%

Tableau 7.5 – Fréquence et proportions des segments en fonction de leur taille

On remarque que plus de 85% des segments contiennent au maximum 5 éléments (*Prop. S. cumul*). Les segments contenant deux éléments forment près de 23%, et ceux en contenant trois, 18%. La majorité des segments se répartissent entre des tailles de 1 à 10.
La distribution du nombre d'éléments dans des segments en fonction de la taille est légèrement différente. Près de 70% des éléments sont contenus dans des segments de taille inférieure à 5, et 95% pour une taille inférieure à 10.

On peut également décrire les segments en fonction de leur type, c'est-à-dire la nature de leurs frontières. Le tableau (7.6) indique les mêmes informations que précédemment vis-à-vis du type : d'une part, le nombre de segment (*NB S.*) et d'autre part le nombre d'éléments concernés (*NB E.*), ainsi que leurs proportions (*Prop.*) et proportions cumulées (*Prop. Cumul.*).

Type	Nb S	Nb. S cumul.	Prop. S.	Prop. S. cumul.	Nb. E	Nb. E. cumul.	Prop. E.	Prop. E. cumul.
,|,	459702	459702	19,05%	19,05%	1163024	1163024	15,43%	15,43%
,|.	281657	741359	11,67%	30,72%	1027209	2190233	13,63%	29,07%
.|,	251479	992838	10,42%	41,14%	703517	2893750	9,34%	38,40%
"|"	101908	1094746	4,22%	45,36%	416932	3310682	5,53%	43,94%
.|.	85170	1179916	3,53%	48,89%	417729	3728411	5,54%	49,48%
(|)	80543	1260459	3,34%	52,23%	276935	4005346	3,68%	53,15%
,|qui	59735	1320194	2,48%	54,70%	150419	4155765	2,00%	55,15%
"|,	57178	1377372	2,37%	57,07%	165839	4321604	2,20%	57,35%
,|(	51759	1429131	2,14%	59,22%	127678	4449282	1,69%	59,05%
,|"	48790	1477921	2,02%	61,24%	156603	4605885	2,08%	61,12%
qui|,	37844	1515765	1,57%	62,81%	127997	4733882	1,70%	62,82%
.|"	37480	1553245	1,55%	64,36%	98686	4832568	1,31%	64,13%
,|que	36387	1589632	1,51%	65,87%	94243	4926811	1,25%	65,38%
qui|.	32734	1622366	1,36%	67,22%	129726	5056537	1,72%	67,10%
que|,	28297	1650663	1,17%	68,40%	92051	5148588	1,22%	68,33%
que|.	27286	1677949	1,13%	69,53%	112694	5261282	1,50%	69,82%
"|.	23241	1701190	0,96%	70,49%	85873	5347155	1,14%	70,96%
:|,	22624	1723814	0,94%	71,43%	55482	5402637	0,74%	71,70%
,|qu'	22527	1746341	0,93%	72,36%	60411	5463048	0,80%	72,50%
.|que	22381	1768722	0,93%	73,29%	70352	5533400	0,93%	73,43%

Tableau 7.6 – Statistiques des segments et des éléments en fonction du type de segment

On observe que la distribution selon le nombre de segments d'un certain type ou le nombre d'éléments qu'ils contiennent est relativement similaire : les 20 types de segments les plus fréquents illustrés représentent près de 73% de la totalité et des éléments. Le type de segment le plus fréquent est <,|,> et couvre 19% des segments concernant 15% des éléments. Il est suivi par deux motifs déséquilibrés <,|.> et <.|,>, puis on retrouve les citations, les phrases situées entre deux points et les parenthétiques. On remarque que les segments dont les types sont des conjonctions *que* et *qui* sont également fréquents : ils correspondent à des relatives ou à des complétives.

On peut enfin chercher des corrélations entre le type et la taille des segments, plus particulièrement le nombre d'éléments dans un type de segments selon une taille donnée. Par exemple, 15% des segments de type <,|,> n'ont qu'un élément et 18% en ont deux. Le tableau (7.7) résume les caractéristiques des 10 segments les plus fréquents en fonction du nombre d'éléments uniquement.

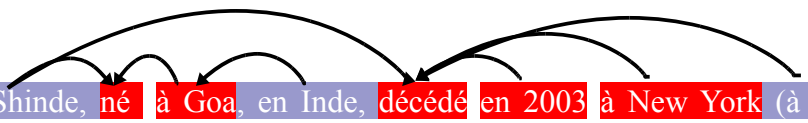
Type	Taille	Nb. E.	Prop. E.	Prop. E. cumul.	Type	Taille	Nb. E.	Prop. E.	Prop. E. cumul.
,|,	2	214646	18,46%	18,46%	(|)	1	184848	66,75%	66,75%
	3	202458	17,41%	35,86%		2	40196	14,51%	81,26%
	1	182284	15,67%	51,54%		3	26145	9,44%	90,70%
	4	163152	14,03%	65,57%		4	10308	3,72%	94,43%
	5	126120	10,84%	76,41%		5	7035	2,54%	96,97%
,|.	4	155556	15,14%	15,14%	,|qui	2	77262	51,36%	51,36%
	3	154941	15,08%	30,23%		3	15084	10,03%	61,39%
	5	138320	13,47%	43,69%		4	14232	9,46%	70,85%
	2	117780	11,47%	55,16%		5	11055	7,35%	78,20%
	6	111396	10,84%	66,00%		6	8124	5,40%	83,60%
.|,	3	102774	14,61%	14,61%	"|,	3	27879	16,81%	16,81%
	2	102654	14,59%	29,20%		4	26276	15,84%	32,66%
	4	100128	14,23%	43,43%		2	25960	15,65%	48,31%
	1	95276	13,54%	56,98%		5	20870	12,58%	60,89%
	5	83800	11,91%	68,89%		1	17900	10,79%	71,69%
"|"	1	116865	28,03%	28,03%	,|(	1	25708	20,14%	20,14%
	2	101124	24,25%	52,28%		3	18852	14,77%	34,90%
	3	57170	13,71%	66,00%		4	17376	13,61%	48,51%
	4	36474	8,75%	74,74%		2	15706	12,30%	60,81%
	5	26915	6,46%	81,20%		5	14180	11,11%	71,92%
.|.	5	55745	13,34%	13,34%	,|"	3	28956	18,49%	18,49%
	6	51990	12,45%	25,79%		4	25752	16,44%	34,93%
	4	51300	12,28%	38,07%		2	22684	14,49%	49,42%
	7	47229	11,31%	49,38%		5	20565	13,13%	62,55%
	8	37312	8,93%	58,31%		6	15042	9,61%	72,16%

Tableau 7.7 – Statistiques des segments en fonction de la taille et du type

Les segments de taille 1 constituent également une proportion significative des citations (28%). Ce fait reflète le comportement particulier des segments entre guillemets : une citation ne correspond pas nécessairement à une phrase, mais peut uniquement contenir un syntagme que l'auteur a souhaité mettre en exergue, pour citer une personne ou simplement pour indiquer une précaution. En observant les occurrences, on constate également que ces guillemets sont une source importante d'erreur de chunking, car ils ne sont pas nécessairement situés à la frontière de syntagmes, mais peuvent séparer un adjectif de son nom ou un auxiliaire de son participe. Il semble donc que le chunking serait plus performant si la gestion des citations était réalisée en amont. Nous n'avons pas pu évaluer cette configuration. La forte présence de parenthétiques de taille 1 (66%) indique que les éléments inclus sont généralement des entités-R ou des chunks. Ils constituent également 13% des segments de type <.|,>, 20% des segments pour le type <,|(> et 10% pour le type <"|,>, qui sont des segments déséquilibrés. En revanche, 58% des segments entre points ont une taille en élément entre 4 et 8.

Cette distribution invite deux perspectives d'analyse différentes : l'analyse intrasegment et l'analyse inter-segment. Les segments ne contenant qu'un élément peuvent être considérés comme des segments « prêts-à-associer » à des éléments des segments adjacents.

Pour illustrer en contexte, nous avons indiqué dans l'exemple (212), les relations (flèches) que l'on pourrait chercher entre les segments de taille 1 (en violet) et les segments plus longs (en rouge).



(212)A. V. Shinde, né à Goa, en Inde, décédé en 2003 à New York (à l' âge de 86 ans), avait parcouru le monde en quête des plus belles pierres d'Orient et d'Occident pour le joaillier H Winston

- L'analyse intra-segment consistera à analyser les relations entre, d'une part *né* et *à Goa* et d'autre part, entre *décédé*, *en 2003* et *à New York*.
- L'analyse inter-segment consistera à identifier les relations entre *A. V. Shinde* et *né*, entre *en Inde* et *à Goa* et entre *à l'âge de 86 ans* et *décédé*.

Un indicateur efficace pour guider l'analyse est la nature des éléments contenus dans chaque segment. Pour obtenir un vue d'ensemble, on peut générer des n-grammes des éléments qu'ils contiennent. Par exemple étant donné que les segments de taille 1 ne contiennent qu'un élément, leur n-gramme est un uni-gramme. Il est possible de réaliser des n-grammes pour chaque niveau de représentation (forme, entité-R, chunk) ou par combinaison de leurs propriétés. Nous avons fait figurer dans le tableau (7.8) les n-grammes les plus fréquents pour les segments de taille 1 à 4 ; ils sont caractérisés en fonction du chunk (lorsqu'il existe) et de l'entité-R tête du chunk.

1-gram	Fréq.	Prop.	3-gram	Fréq.	Prop.
_pers	59216	7,83%	*_pronom|_action|GN_subs*	16045	4,34%
GN_subs	55565	7,35%	*_pronom|_aux|GN_subs*	11503	3,11%
_val_score	54169	7,16%	*_pronom|_pronom|_action*	8016	2,17%
_adjectif	51693	6,84%	*_pronom|GV_action|GN_subs*	7191	1,94%
_subs	42232	5,59%	*_action|GN_subs|GNP_de_subs*	5452	1,47%
_time	33863	4,48%	*GN_subs|_action|GN_subs*	5316	1,44%
_org	28987	3,83%	*_pronom|_action|_adv*	5072	1,37%
_loc	28748	3,80%	*_pers_bof|_pronom|_action*	4280	1,16%
_action	28496	3,77%	*_np|_punct|_acro_div*	3992	1,08%
_np	26824	3,55%	*_pronom|_action|_action*	3961	1,07%
Total	409793	54,19%	**Total**	70828	19,15%
2-gram	**Fréq.**	**Prop.**	**4-gram**	**Fréq.**	**Prop.**
_pronom|_action	36992	6,57%	*_pronom|_action|GN_subs|GNP_de_subs*	3147	1,23%
_action|GN_subs	23635	4,20%	*_pronom|_aux|GN_subs|GNP_de_subs*	2792	1,09%
GN_subs|GNP_de_subs	17576	3,12%	*_pronom|_action|_action|GN_subs*	2528	0,99%
_pronom|GV_action	14370	2,55%	*_pronom|_action|_adv|GN_subs*	1888	0,74%
_action|_pers	13321	2,36%	*_pronom|GV_action|GN_subs|GNP_de_subs*	1817	0,71%
_action|_pronom	11734	2,08%	*GN_subs|_action|GN_subs|GNP_de_subs*	1414	0,55%
GN_subs|_action	11229	1,99%	*_pronom|_action|_pronom|_action*	1372	0,54%
_pronom|GVADJ_adjectif	9844	1,75%	*_pronom|GV-NEG_action|_action|GN_subs*	1342	0,53%
_prod|GNP_du_time	9142	1,62%	*_pronom|_pronom|_action|GN_subs*	1280	0,50%
_subs|GNP_de_subs	9058	1,61%	*_pronom|_aux|_adv|GN_subs*	1274	0,50%
Total	156901	27,85%	**Total**	18854	7,39%

Tableau 7.8 – N-grammes les plus fréquemment observés dans les segments

On remarquera tout d'abord que le nombre d'occurrences couvertes par n-gramme décroit en fonction de sa taille : 54% des unigrams sont couverts par 10 d'entre eux alors que les 10 4-grams les plus fréquents ne représentent que 7%.

- Concernant les unigrams, on constate qu'une grande portion sont des EN classiques (Organisation, Personne, Lieu), des Npr, ainsi que des des catégories sémantiques de Temps et de Valeur : des catégories informationnelles qui nous intéressent.
- Les bigrams peuvent être distingués en deux groupes : des séquences qui s'apparentent à des syntagmes (*GN_subs|GNP_de_subs*) ou des séquences qui s'apparentent à des sujets (*_pronom*, *GN_subs*, *_pers*) accompagnés de leurs verbes (*_action*), qui peuvent être inversées (*_action|_pers*).

- Les verbes et auxiliaires apparaissent dans la majorité des trigrames et dans tous les 4-grams. Il semble donc que les verbes occupent une place prépondérante dans les segments de taille supérieure à 1.

Enfin, on peut également observer que certaines séquences contiennent des signes de ponctuation.
La segmentation offre un cadre d'analyse structuré et linguistiquement motivé que nous allons utiliser dans deux applications :

- L'analyse de relations inter-segment pour la détection de relation à longue distance (chapitre 8).
- L'analyse de relations intra-segment pour la désambiguisation d'Entités Nommées (chapitre 10).

8. Relations inter-segment

Dans ce chapitre nous nous intéresserons aux relations entre segments dans le corpus de développement (presse) dont nous disposons. Nous présenterons tout d'abord les principes de la grammaire de segments, puis les structures que nous avons étudiées. Les règles définissent des contraintes sur des segments adjacents ainsi que des opérations d'association. La représentation arborescente est très efficace pour illustrer la représentation obtenue, nous en ferons donc largement usage. L'application de ces règles résulte en une restructuration de l'arbre d'un document et modifie la distribution des segments.

8.1. Principes de la grammaire de segment

Comme nous l'avons noté, le nombre de segments simples (de taille 1) constitue plus de 30% de la totalité des segments du corpus de développement. Nous reproduisons en figure (8.1) plus en détail la nature des entités-R apparaissant dans ce contexte.

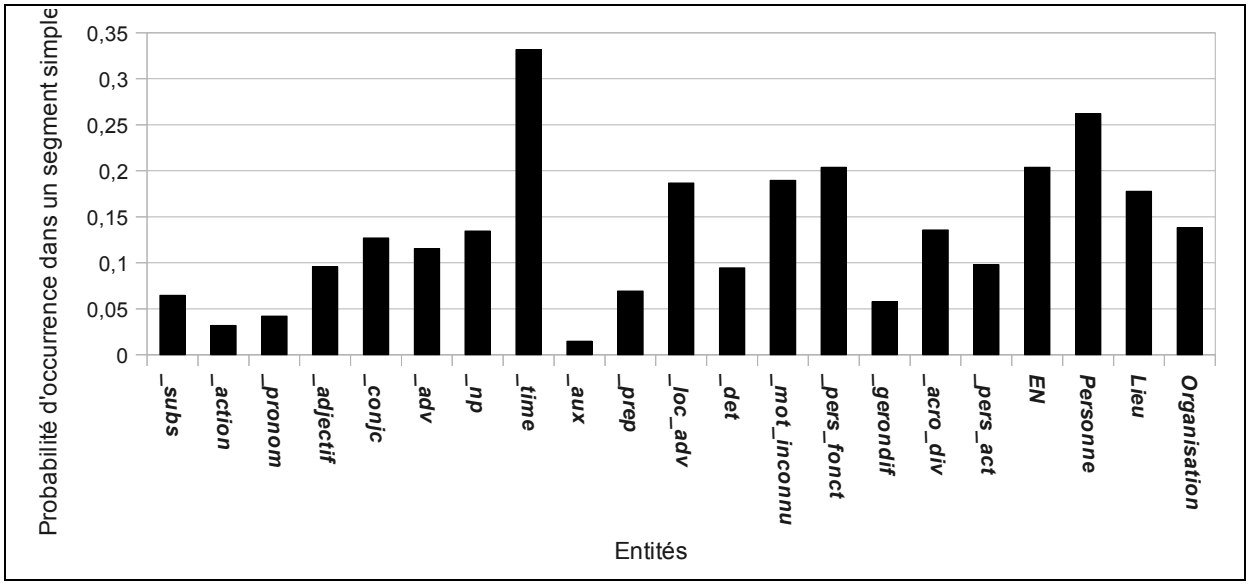


FIG 8.1 – Probabilité d'occurrence d'une entité-R dans un segment simple

On observe qu'environ 21% des occurrences d'EN (Organisation, Personne, Lieu) détectées par le système Rnc se retrouvent dans des segments simples (catégorie EN), c'est-à-dire qu'une frontière se situe à leur gauche et à leur droite. À titre comparatif, les entités-R se retrouvant fréquemment dans cette configuration sont des syntagmes temporels (0,33), des locutions adverbiales et des mots inconnus (0,18), ainsi que des entités-R de fonction (0,21).
Nous avons également fait figurer la répartition des EN selon la catégorie et on observe que l'on trouvera avec plus de certitude une personne (0,26) qu'une organisation (0,14) dans un segment simple. L'intérêt du point de vue des EN est d'identifier les relations qui les lient aux éléments de segments adjacents.

Pour illustrer le principe de rattachement des segments, et du fonctionnement général de la grammaire de segments, prenons un exemple (213).

(213)Le sommet des Amériques, réuni à Monterrey (Mexique), les 12 et 13 janvier, se penche sur la "gouvernance démocratique" et sur les incidences de la corruption .

D'une part, on remarque que cette phrase contenant 7 segments est composée de trois EN, deux de type Lieu (*Monterrey*, *Mexique*) et une de type événement (*sommet des Amériques*) et que chacune figure dans un segment différent. D'autre part, on observe que le sujet est séparé de son verbe par 3 segments. La figure (8.2) montre les arcs de dépendance entre segments.

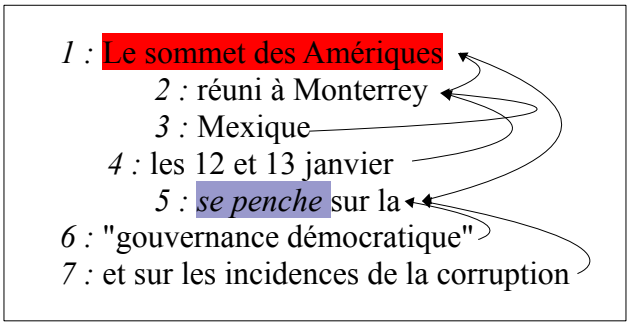


FIG 8.2 – Distances et dépendances entre segments

À partir de l'analyse des arcs de dépendance, on peut identifier l'ordre dans lequel doivent s'effectuer des règles inter-segment portant sur des segments adjacents. Par exemple, le segment 4 ne pourra être associé au segment 2 qu'après que le segment 3 ait été associé au segment 2. L'exemple (214) traduit la suite des opérations possible pour parvenir à détecter le sujet.

(214) 1= (Mexique)
2= , réuni à Monterrey (Mexique)
3= , réuni à Monterrey (Mexique) , les 12 et 13 janvier ,
4= Le sommet des Amériques , réuni à Monterrey (Mexique) , les 12 et 13 janvier ,
5= Le sommet des Amériques , réuni à Monterrey (Mexique) , les 12 et 13 janvier , se penche sur la

L'état 1 est le point de départ. La structure que rend le segmenteur est illustrée figure (8.3).

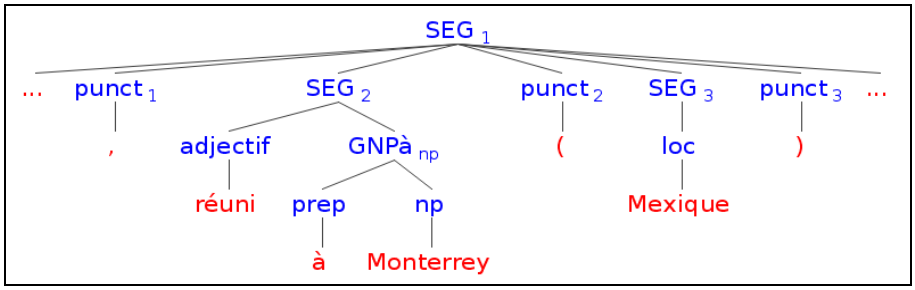


FIG 8.3 – Arbre fourni par le segmenteur pour l'exemple (214)

Dans ce cas, l'EN *Mexique* précise sémantiquement la localisation de la ville *Monterrey* (qui n'est par ailleurs pas reconnue comme lieu) : on cherche donc à la rattacher, non pas au segment précédent, mais à cet élément précis. Cette opération nécessite de déplacer tous les éléments du second segment (la parenthétique) ainsi que ses frontières sous la dépendance de *Monterrey*. Le résultat attendu est illustré figure (8.4).

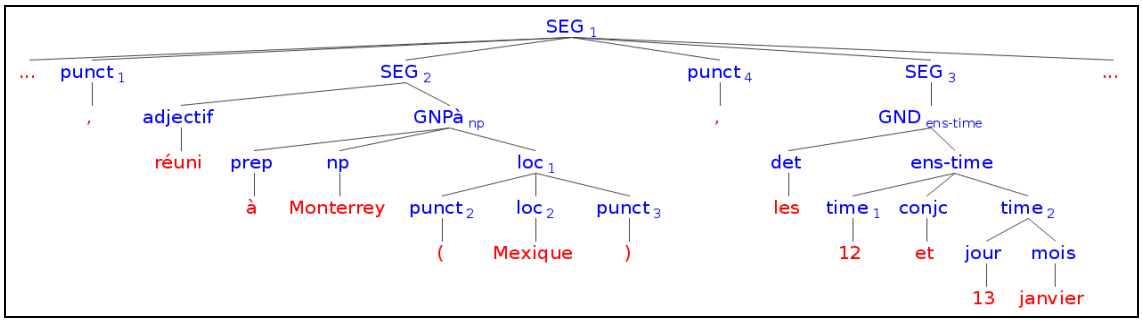


FIG 8.4 – Arbre pour l'exemple (214) après association de segments

Cette représentation permet de traduire correctement la dépendance entre l'élément contenu dans la parenthèse et l'élément du segment qui le précède. Les opérations qu'une telle association suppose sont :

1. la suppression du nœud du segment 3
2. la création d'un nouveau nœud
3. le positionnement de ce nœud sous l'élément-cible (chunk ou entité-R) du segment 2
4. l'inclusion des parenthèses et des éléments du segment 3 supprimé dans ce nouveau nœud
5. la catégorisation de ce nœud en fonction de l'élément principal qu'il contient (ici le lieu)

Cette dernière opération doit permettre d'extraire facilement la relation entre *Monterrey* et *Mexique*.

L'application de cette règle permet d'alléger l'arbre et de rapprocher d'autres segments. On peut donc analyser les relations du segment 2 « augmenté » avec le segment 4 *les 12 et 13 janvier*. Dans ce cas, les mêmes opérations ne peuvent être répliquées, car l'apposition temporelle est en relation avec l'ensemble du segment 2 dont la tête est le participe *réuni*. Il s'agit d'un problème de positionnement du nouveau nœud. La représentation attendue est illustrée dans la figure (8.5).

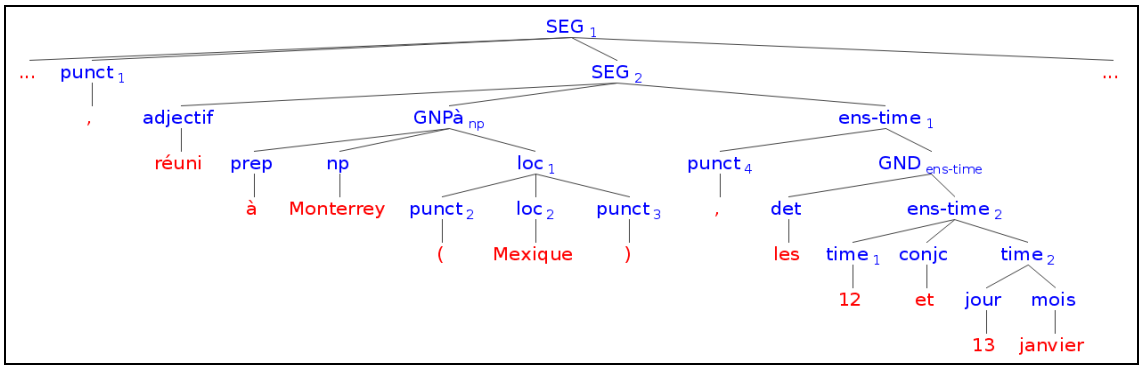


FIG 8.5 – Arbre pour l'exemple (214) après association de segments

Les deux opérations restantes consisteront à associer le segment 2 à la tête du segment 1 (*le sommet des Amériques*) et à fusionner le segment 1 avec le segment 5 (contenant le verbe principal), ce qui permettra d'obtenir l'arbre illustré figure (8.6).

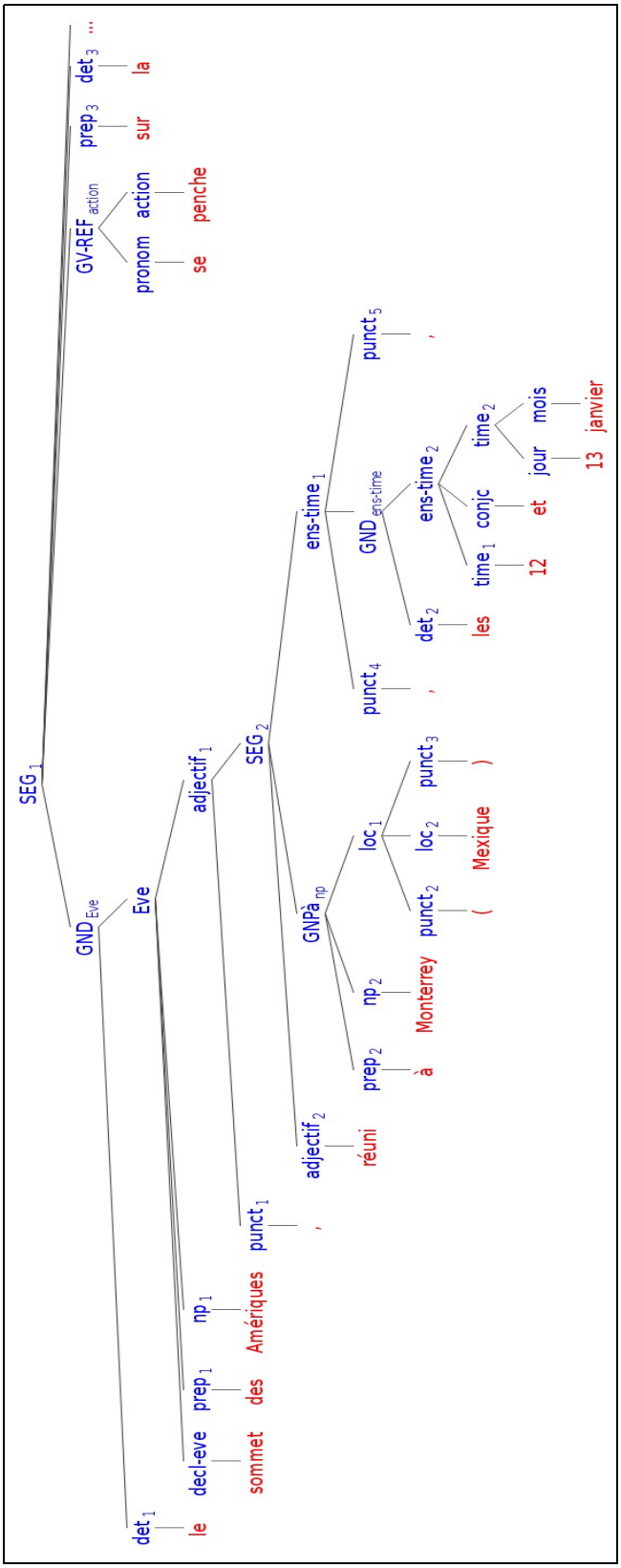


FIG 8.6 – Arbre pour l'exemple (214) en fin d'association de segments

Réaliser une grammaire de segments permet de restructurer l'arbre d'un article ou d'une phrase en associant les segments deux à deux. L'identification de ces relations nous mène tout naturellement à une fusion des segments (faibles ou semi-faibles), ce qui résulte en une simplification progressive de la structure de la phrase. Comme nous pouvons l'observer, une des conséquences de cette fusion est la réduction de la distance entre des expressions linguistiques liées par dépendance syntaxique. Dans la figure (8.6), le GN ayant pour tête l'entité-R *_Eve* (événement) est le nœud sœur du groupe verbal pronominal (*GV-REF*).

Plutôt que de parcourir l'espace gauche du verbe pour identifier le sujet syntaxique comme on pourrait le faire en concevant un analyseur syntaxique symbolique, cette méthode permet d'élaguer l'arbre et de faire émerger le sujet directement à gauche du verbe. Notre méthode partage des similarités avec la tâche de compression de phrase en TAL, étudiée dans le cadre du résumé automatique de textes [Yousfi-Monod, 2007]. Nous n'avons néanmoins pas connaissance de travaux alliant cette problématique avec celle des EN.

8.2. Détection de relation entre segments

Les règles que nous avons développées sont au nombre de 28 et s'appuient sur la catégorie sémantique des entités-R, la nature des segments et leur taille. Elles sont classées en quatre types que nous allons à présent décrire. Elles ont été développées grâce au système décrit chapitre 9. La description du fonctionnement du système n'est pas essentiel pour aborder le sujet des relations inter-segment, mais on pourra s'y reporter pour plus de détail.

8.2.1. Les parenthétiques

La délimitation des parenthétiques est effectuée par le segmenteur et leur identification dans une règle constitue une tâche relativement aisée. Ce sont des segments dont les frontières initiale et finale sont des parenthèses. Cette règle est prioritaire, car la parenthétique peut être imbriquée dans des appositions qui doivent être traitées ultérieurement. Le rattachement (l'action de la règle) en revanche peut poser certains problèmes comme lorsque l'élément du segment précédant la parenthétique n'est pas correctement isolé dans un chunk. L'algorithme utilisé consiste à inclure la parenthétique sous le dernier nœud à droite du segment précédent. Par exemple, dans une séquence comme (215), on obtient l'arbre illustré figure (8.7).

(215)… des Montoneros (la guérilla péroniste) …

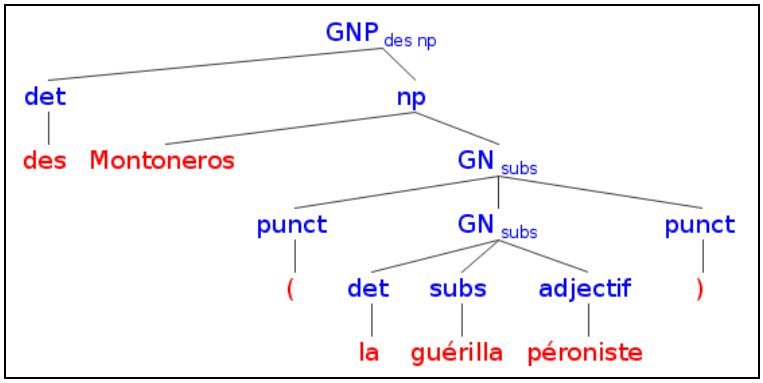


FIG 8.7 – Arbre de l'exemple (215) après association de parenthétique

Le problème principal, pour le rattachement de segment et pour les parenthétiques en particulier, est de pouvoir identifier le nœud dont ils dépendent : ce peut être le nœud final de plus haut niveau, un nœud intermédiaire sous ce nœud ou encore un nœud précédent. Par exemple, certaines entités-R comme *_pers_fonct* (*président*, etc.), incluent parfois une entité-R de type *Organisation*, comme illustré dans la figure (8.8), où *Sceaux* est l'organisation et *Hauts - de - Seine*, sa localisation.

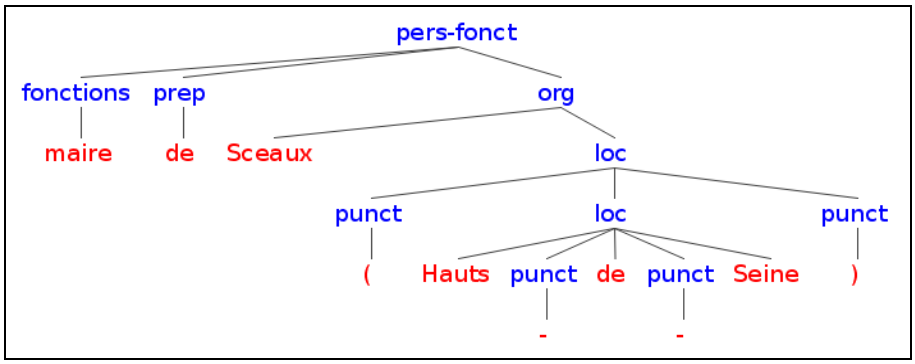


FIG 8.8 – Arbre associant une parenthétique contenant un Lieu

Ce rattachement est préférable à un autre qui placerait la parenthétique sous le nœud de type *_pers_fonct* ou encore à droite de ce nœud puisqu'il explicite la relation entre ces deux entités-R. La décision de rattachement dépend donc à la fois de la nature du nœud inclus dans la parenthétique et des nœuds inclus dans le segment qui la précède. Or le contenu d'une parenthétique peut être extrêmement variable, ce qui invite à une étude plus approfondie et systématique de ces structures dans les corpus écrits : la figure (8.7) était un cas de relation de reformulation-définition, alors que la figure (8.8) illustre un cas de relation de localisation. Les parenthétiques renferment généralement des informations pertinentes sur les EN qui les précèdent, mais la nature de la relation peut être très variée. Nous avons identifié 12 relations majeures indiquées dans le tableau (8.1).

Relation	Exemple
Localisation	*à Scheveningen (Pays - Bas)*
Dénomination complète d'un acronyme	*les ASF (Autoroutes du sud de la France)*
Acronyme d'une dénomination	*le Parti révolutionnaire institutionnel (PRI)*
Score d'un événement de type compétition sportive	*à l' issue de la nette victoire (85 - 65)*
Date d'un événement	*l' élection du président Ernesto Samper (1994)*
Organisation dont une personne est membre	*José Serra (Parti social - démocrate brésilien)*
Dates de naissance et de mort	*le roi Louis VI (1108 - 1137)*
Relation Type – EN	*deux délégations (CGPME et UPA)*
Relation EN – Type	*l' entreprise Maugein (la plus ancienne marque d' accordéons de France)*
Relations entre valeurs (absolu/relatif)	*un chiffre d' affaire de 5 millions d' euros (+ 13 pourcent)*
Relation typée par la parenthétique	*la convention Unedic (signée par le Medef, la CFDT, la CFE - CGC et la CFTC)*
Source inter-textuelle	*regrettait récemment M. Almunia (Le Monde du 10 septembre)*

Tableau 8.1 – Relations sémantiques contenues dans des parenthétiques

Comme on peut l'observer, les parenthétiques contiennent des informations pertinentes sur les EN qui pourraient aider à leur classification.

8.2.2. Les insertions

Ce que nous regroupons sous le nom d'insertion concerne des unités fortement diversifiées dont la présence crée une rupture de l'ordre syntagmatique canonique d'une phrase. Il s'agit d'unités apposées qui séparent un groupe verbal d'un groupe nominal et dont les frontières sont des virgules. Elles jouent parfois un rôle identique aux parenthétiques, notamment dans les relations Type – EN et EN – Type. Nous avons limité leur détection en fonction du type d'entité-R de l'élément initial de l'insertion et en fonction du type d'entité-R finale du segment précédent. Elles sont de quatre types : les relations EN – Type, les relations Type – EN, les insertions adverbiales et les participiales.

1. La relation EN – Type : les EN retenues sont les lieux les organisations et les personnes et sont situées dans des segments ne contenant pas de verbe. Les Types retenus pour l'EN *Personne* sont les fonctions (*président*, etc.) ou activités (*chanteur*, etc.), les superlatifs ou comparatifs (*premier*, etc.), les entités-R d'origine (français, etc.), de famille (*père*, etc.) ou d'âge (*34 ans*, etc.). Pour les lieux et organisations, nous avons restreint les types aux superlatifs et aux types d'organisation (*commission*, etc.) en ajoutant les adresses pour les lieux.
2. La relation Type – EN emploie essentiellement les mêmes critères, mais la position des segments est inversée et le type de chunk est contraint.

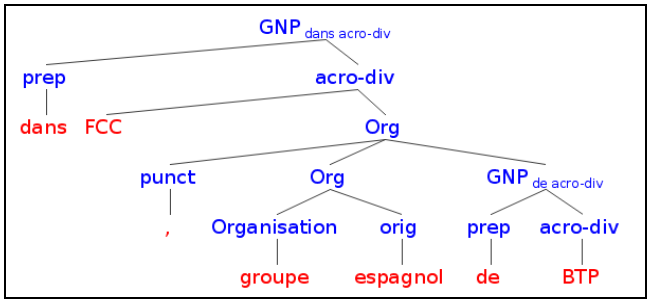


FIG 8.9 – Relation Organisation-Type

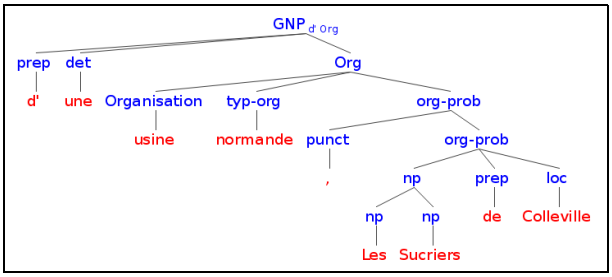


FIG 8.10 – Relation Type-Organisation

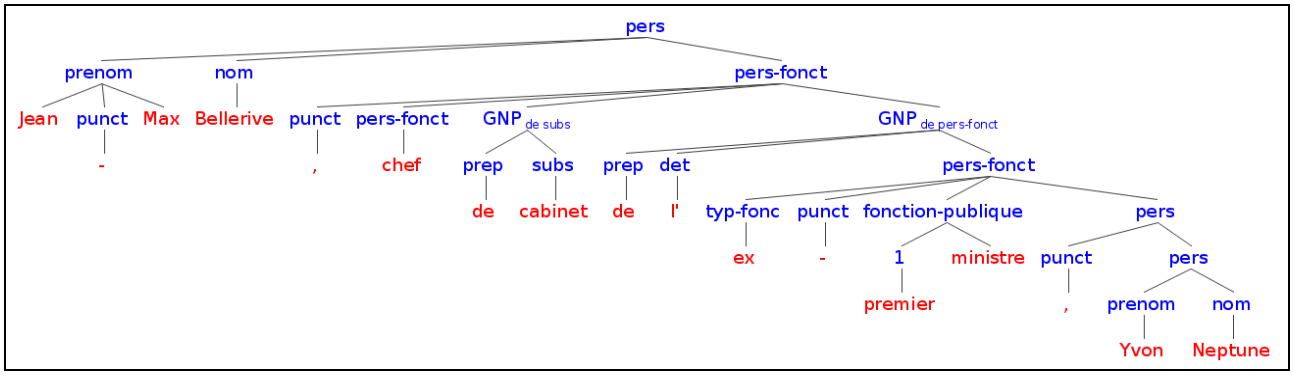


FIG 8.11 – Relations Personne-Type et Type-Personne

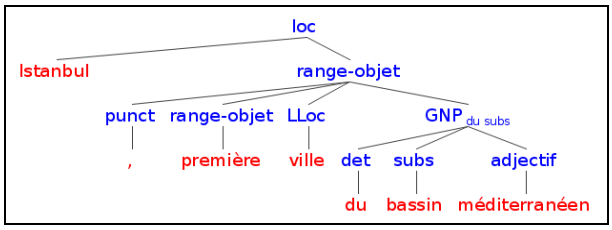


FIG 8.12 – Relation Lieu-Type

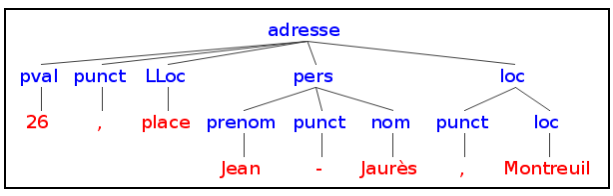


FIG 8.13 – Relation Type-Lieu

Pour les relations EN – Type, nous avons également sélectionné les adjectifs (figure 8.14) et les substantifs (figure 8.15) non inclus dans des chunks, qui désignent des EN de type fonction non identifiées comme telles par le système Rnc.

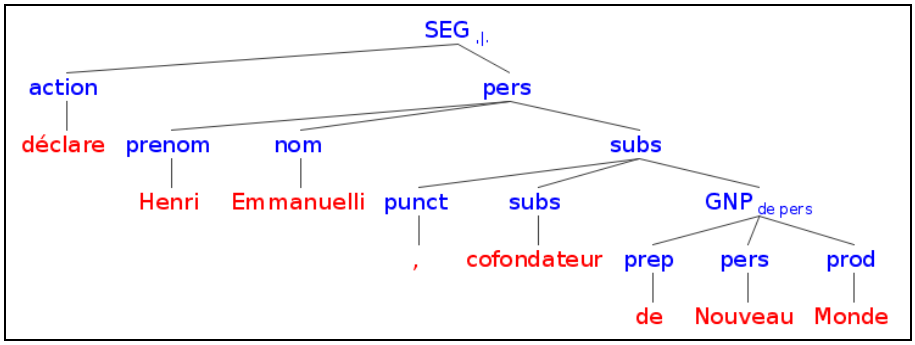


FIG 8.14 – Arbre après association d'insertion dont l'entité-R initiale est un substantif

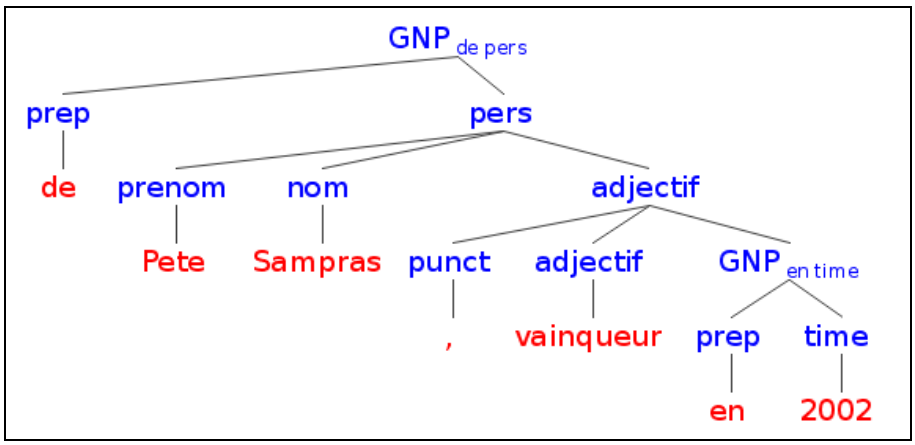


FIG 8.15 – Arbre après association d'insertion dont l'entité-R initiale est un adjectif

3. L'insertion adverbiale : de nombreuses entités-R adverbiales gravitent autour des EN et des verbes. C'est notamment le cas des groupes prépositionnels (*de son côté*, *selon X*, etc.), des entités-R temporelles (*en 1998*, etc.) et des locutions adverbiales (*bien sûr*, etc.). Selon les cas, elles peuvent dépendre de l'EN ou du verbe. Ces insertions peuvent précéder (figure 8.16) ou suivre (figure 8.17) ces éléments.

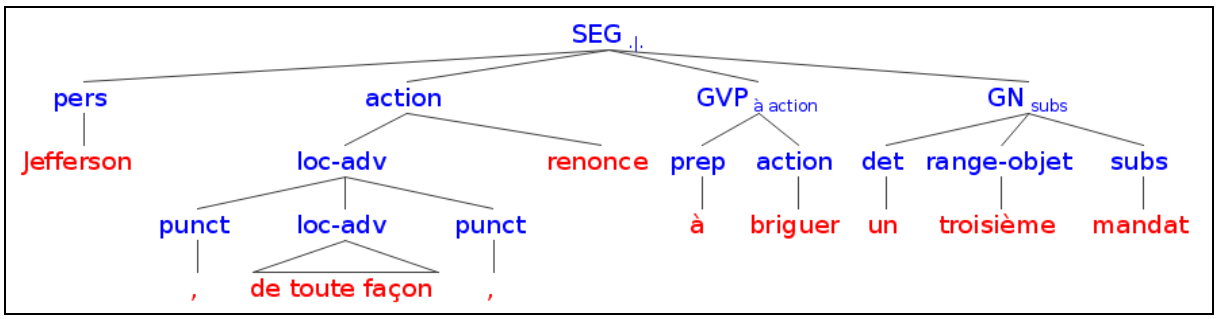


FIG 8.16 – Arbre obtenu par association d'insertion contenant une locution adverbiale

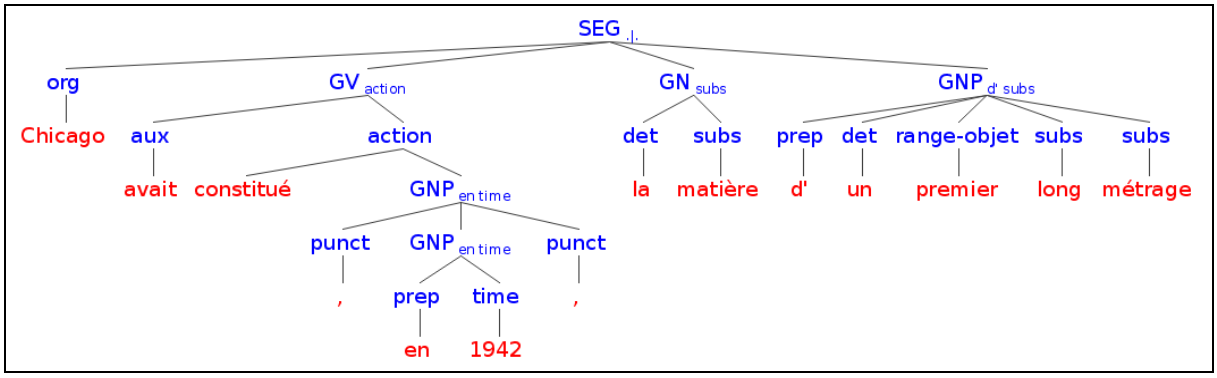


FIG 8.17 – Arbre obtenu par association d'insertion contenant une expression temporelle

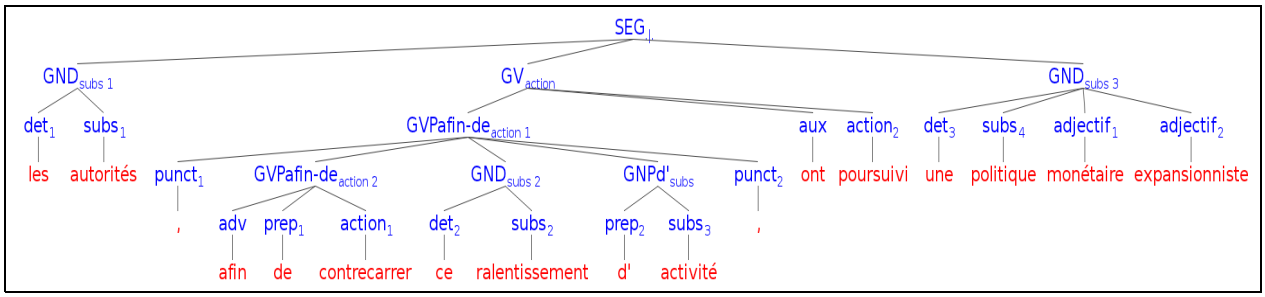


FIG 8.18 – Arbre obtenu par association d'insertion contenant un groupe prépostionnel

4. Les participiales : elles sont proches des insertions EN – Type mais la nature de la relation est plus variable. La détection des participes passés est effectuée au moyen d'une liste lexicale extraite du lexique Morphalou[46].

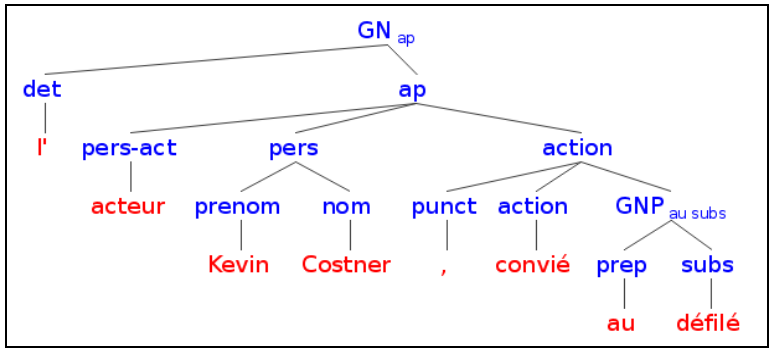


FIG 8.19 – Arbre obtenu par association d'insertion de participiale

46 http://www.cnrtl.fr/lexiques/morphalou/

8.2.3. Les listes

Les listes (ou énumérations) sont des structures dans lesquelles apparaissent régulièrement les EN. Comme l'ont noté L. Kosseim et. T. Poibeau [Kosseim & Poibeau, 2001], ces structures ont un potentiel de désambiguïsation. La différence entre une insertion et une liste repose sur la dimension d'équilibre : les entités situées entre virgules sont de même catégorie. Une liste ne répond alors pas au schéma de dépendance syntaxique classique en tête/gouverneur puisque toutes les EN sont considérées sur le même plan. Pour représenter cette propriété, il faut les associer au même niveau hiérarchique.

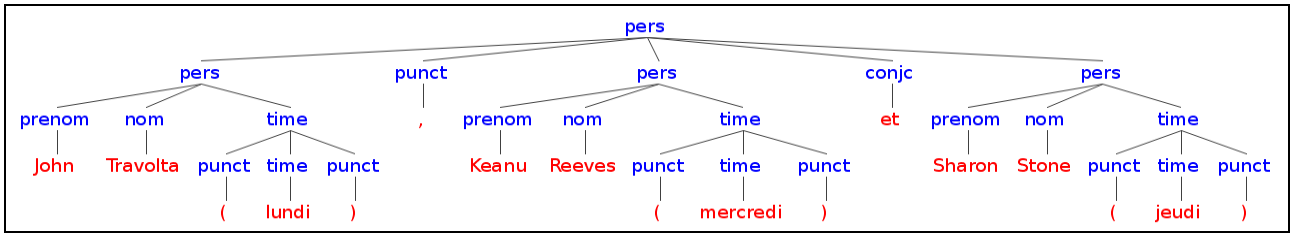


FIG 8.20 – Arbre obtenu par association de listes de personnes

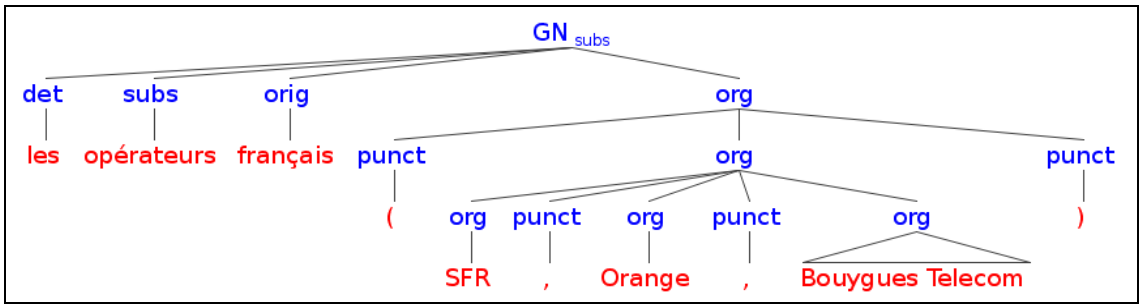


FIG 8.21 – Arbre obtenu par association de listes d'organisations

Les listes que nous traitons sont limitées à des segments associés de taille 1, mais ne sont pas limitées aux EN. Une première règle associe des éléments d'entité-R et de chunk identiques (les prépositions sont lemmatisées), ce qui permet également de traiter des listes de verbes ou de substantifs. Nous avons également développé quelques heuristiques pour prendre les cas de coordination ou de variation faible de catégorie (équivalences adjectif/verbe ou EN de différente nature) en compte (figure 8.22).

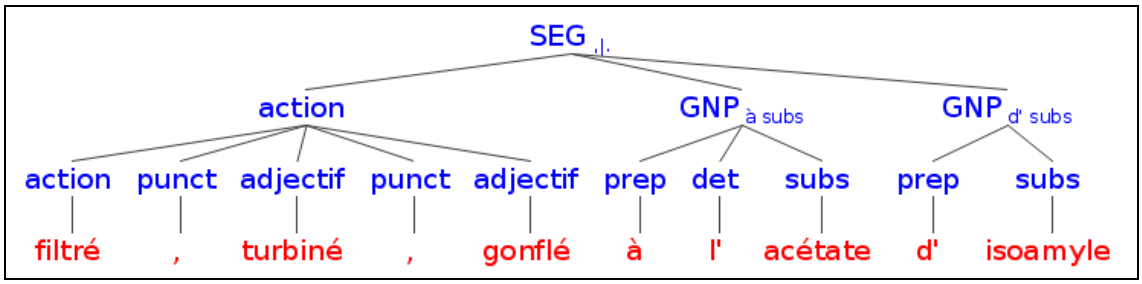


FIG 8.22 – Arbre obtenu par association de listes de verbes

8.2.4. Les relatives

Les relatives sont des propositions syntaxiquement cohérentes. Elles contiennent un ou plusieurs verbes et sont initiés par un pronom relatif. Contrairement à d'autres propositions, elles dépendent d'un groupe lié au verbe de la relative par une relation syntaxique (son sujet, son objet, etc.). Par exemple en (216), la relative dont le verbe est *battre* dépend syntaxiquement du chunk nominal de type *Personne* dont la tête est *Shang Shichun*.

(216)Le Chinois Shang Shichun, qui avait battu trois records du monde dans la catégorie des 75 kg
On peut donc représenter cette dépendance en plaçant la relative sous l'entité-R qui en est tête, comme illustré dans la figure (8.23).

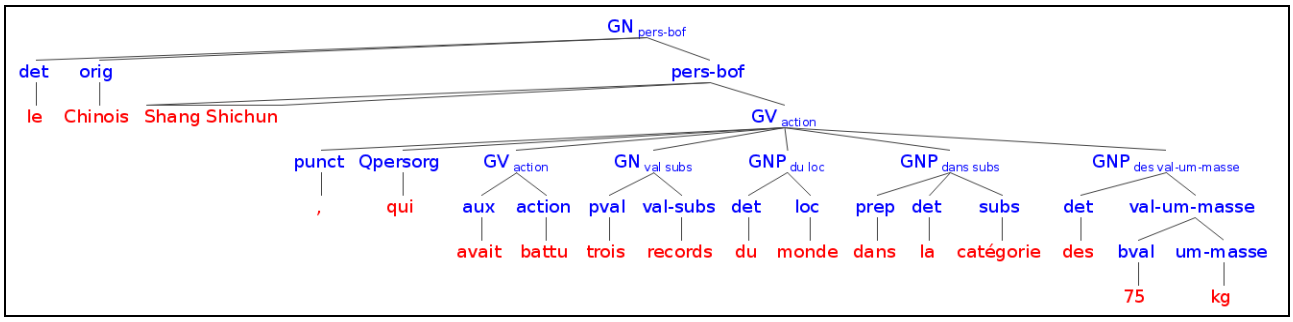


FIG 8.23 – Arbre obtenu par association de relative (exemple 216)

8.2.5. Limites

Le rattachement des relatives montre une des limites de notre méthode, car si dans l'exemple précédent, la relative est appositive (optionnelle, située entre virgules), la frontière droite d'un segment initié par un pronom relatif peut dépasser la portée de la proposition relative. En déplaçant la relative, on risque ainsi de déplacer des éléments qui n'en font pas partie. Le déplacement de la relative restrictive en (217) résultera en une mauvaise segmentation (figure 8.24).

(217)Les gamins qui shootent dans un filet orange bourré de paille vivent une enfance bien différente.

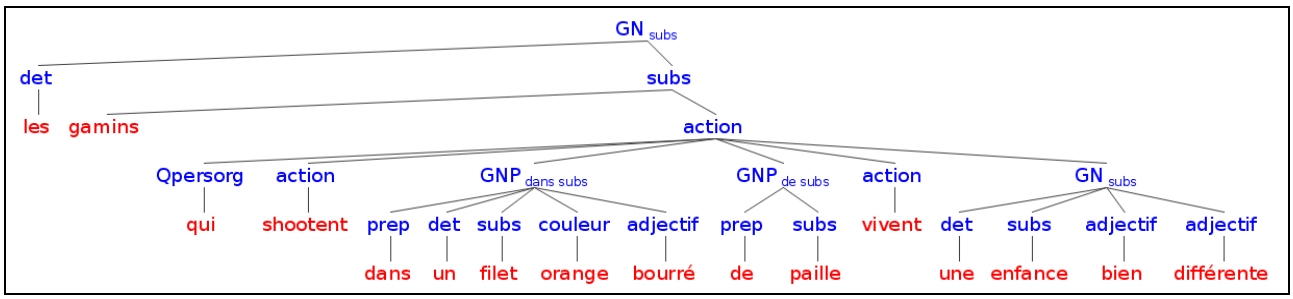


FIG 8.24 – Arbre obtenu par association de relative

On pourrait segmenter en utilisant le verbe comme frontière avant ou après l'application de ces règles. Dans le premier cas, cela signifierait obtenir des données extrêmement fragmentées ; dans le second cas, cela signifierait de ne segmenter que dans les relatives. Nous ne considérons pas la possibilité d'utiliser une analyse syntaxique, bien qu'on puisse tout-à-fait concevoir des modèles hybrides. Les relatives sont limitées aux segments initiés par le pronom *qui* (nous incluons également le pronom *où*). La détection de relatives en *que*, suppose qu'on les distingue des complétives qui sont des propositions autonomes. Nous ne les traitons pas.

Ces phénomènes discursifs de surface ne constituent qu'une faible partie des relations existant entre segments. Par exemple, nous n'avons pas traité les incises qui seraient difficilement représentables dans un DAG (« Directed Acyclic Graph » ou graphe orienté acyclique ; un arbre), étant incluses dans un segment (218).

(218)c'est quelqu'un, dit - elle , qui a beaucoup compté pour moi

Le comportement par défaut serait de l'associer au dernier élément du segment gauche pour permettre de poursuivre l'analyse. Il serait possible d'assigner à ces incises une catégorie spéciale qui permette de les retrouver dans l'arbre.

Comme les règles inter-segments reposent sur les catégories identifiées préalablement par LoRit et Rnc, elles sont soumises à la qualité de leur annotation : par exemple, les erreurs de la grammaire inter-segment nous ont permis d'identifier des mauvais rattachements du chunking, pouvant être dus à une erreur d'annotation du système Rnc. En (219), par exemple la frontière de l'EN est mal détectée : *Marriott Management* et *Services* font partie de deux nœuds séparés. La règle de segmentation associe alors le segment entre virgules à *Services* (figure 8.25).

(219)il a été l' artisan du rachat de Marriott Management Services, numéro un américain de la restauration collective

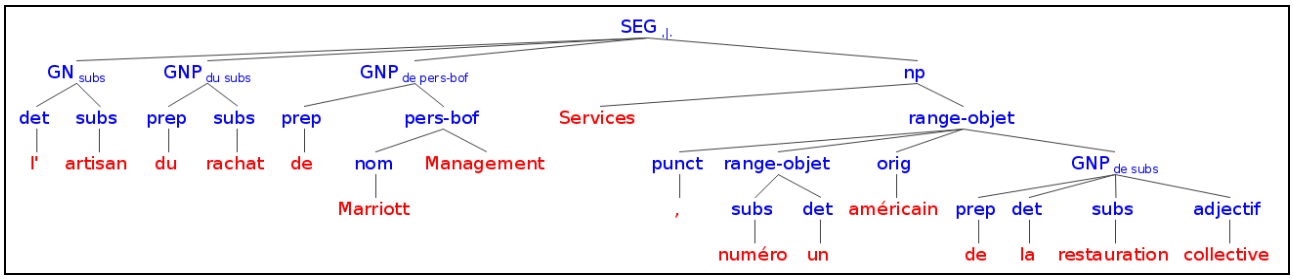


FIG 8.25 – Arbre obtenu pour l'exemple (219)

On peut évidemment imaginer qu'une deuxième phase de traitement des EN corrige ces erreurs de détection. Cela impliquera de positionner *Services* comme nœud sœur de *Management* et de faire remonter l'insertion d'un niveau, autrement dit de fusionner les nœuds *pers-bof* et *np*.

Il faut enfin préciser que les opérations de rattachement de segments deviennent de plus en plus complexes au fur et à mesure qu'ils s'imbriquent : il faut pouvoir prévoir les différents cas de figure et ordonner leur application. Une virgule non déplacée peut occasionner un blocage des futures règles. Il n'y a pas de recette miracle : les opérations de rattachement sont effectuées par de nombreux essais-erreurs à partir de corpus. Ce qui nous ramène à la question de la spécificité des règles inter-segments vis-à-vis du style des articles. Notre grammaire s'applique aux articles de presse, mais elles pourraient probablement s'appliquer à d'autres styles ou domaines. Pour le savoir,

il faudrait pouvoir comparer les résultats obtenus.

Nous nous sommes limité aux segments situés au sein d'une proposition, mais on peut envisager l'étude de relations discursives comme celles qui sont proposées par la RST (Rhethorical Structure Theory ; [Mann & Thompson, 1988]) pour guider l'analyse de segments au-delà des propositions.

Une des propriétés intéressantes de cette méthode d'association de segments est qu'elle fait émerger les groupes d'information principaux, en élaguant les diverses formes d'insertions. C'est cette propriété que nous évaluons dans le prochain chapitre.

9. Apport d'une grammaire de segment pour la détection de sujets à longue distance

Notre recherche sur la segmentation discursive était initiée par le constat des difficultés posées par la variation de surface à la détection des relations sémantiques entre verbes et EN. Le formalisme de patrons lexico-syntaxiques rendait complexe la construction de patron, car il supposait d'intégrer de nombreuses alternatives pour chaque relation. À présent que nous avons défini une grammaire de segments qui structure l'arbre en élaguant les éléments optionnels, nous pouvons reprendre ce problème. Pour illustrer son apport, nous proposons de nous intéresser à la relation sujet, lorsque la catégorie est une EN mais aussi dans les cas où il s'agit d'autres catégories. L'objectif est d'évaluer le nombre de relations sujet obtenu avant et après l'application des règles inter-segment. Notre hypothèse est que ces règles permettent de détecter des relations à longue distance : nous devrions donc en obtenir plus. Nous y parviendrons en comparant deux représentations du texte à une même grammaire syntaxique, ce qui nous permettra de mesurer cet apport pour chaque catégorie sujet. Après avoir observé quelques statistiques pertinentes vis-à-vis des relations sujet, la méthode utilisée pour définir la grammaire syntaxique sera présentée. L'évaluation portera sur les différences de performance de cette grammaire à l'extraction des sujets du verbe *annoncer* selon la représentation.

9.1. Nouvelles données

9.1.1. Taille des segments

La grammaire de segment a permis de réduire de 40% le nombre de segments total, pour une moyenne de 40 segments par articles (il était de 70-80). Si on compare la distribution du nombre de segments en fonction de la taille avant et après l'application de la grammaire de segments, on constate que la proportion du nombre de segments de taille 1 et 2 a fortement diminué (figure 9.1).

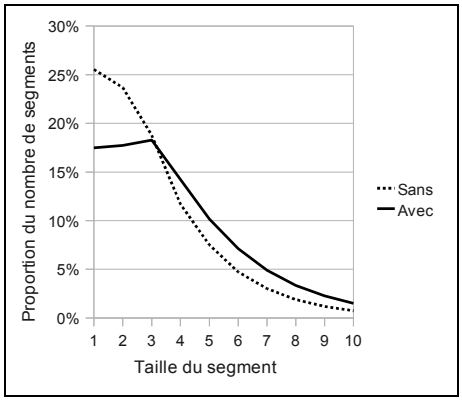


FIG 9.1 – Nombre de segments selon la taille avec/sans application de règles de segments

Cela s'explique principalement par le fait que le contenu des segments courts a été associé et que ces segments ont été supprimés. Si l'on s'intéresse à présent au contenu de ces segments, on peut comparer les segments contenant au moins une EN et les segments contenant au moins un verbe avant et après l'application de la grammaire de segmentation (figure 9.2).

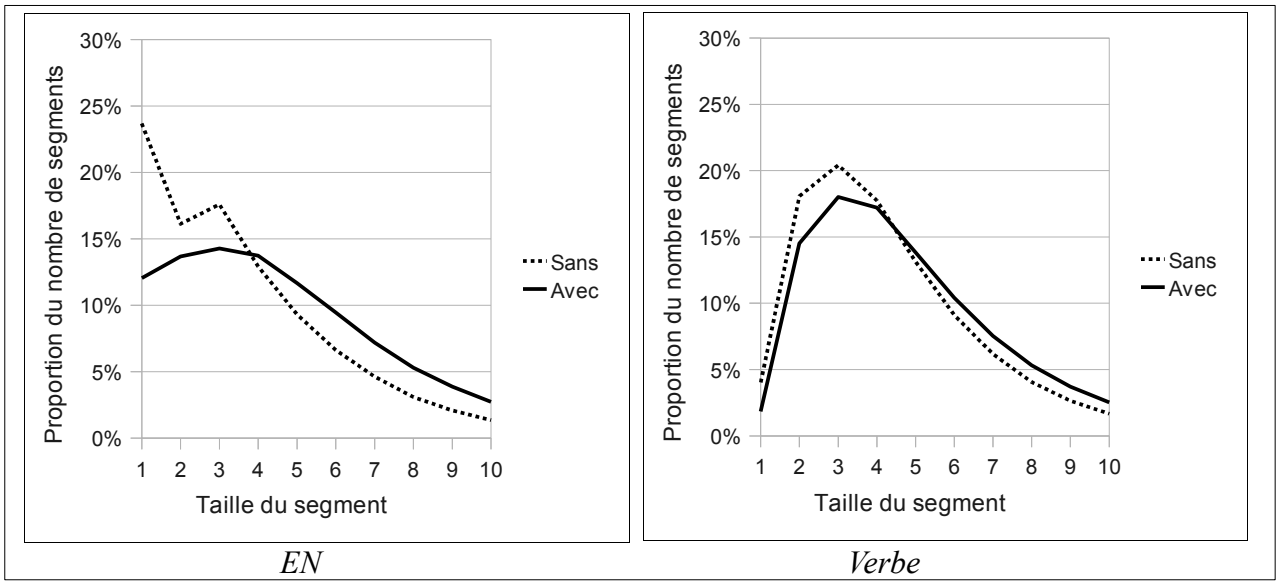


FIG 9.2 – Nombre de Segments contenant une EN (à gauche) et contenant un verbe (à droite) selon la taille, avant/après application de règles de segments

Les EN s'opposent aux verbes sur les segments de taille 1 : la probabilité de trouver une action dans un tel segment est très faible comparée à celle d'y trouver une EN. Les règles de segments ont donc plus d'impact sur la catégorie des EN, réduisant la part de segments de taille 1 de moitié (plus grand écart observé). Après segmentation, le nombre de segments est constamment plus faible pour ces deux catégories jusqu'à la taille 4, à partir de laquelle la part de segments devient plus importante. Les diagrammes suivants précisent ces données pour les sous classes d'EN, *Personne*, *Organisation* et *Lieu* avant et après segmentation (Figure 9.3).

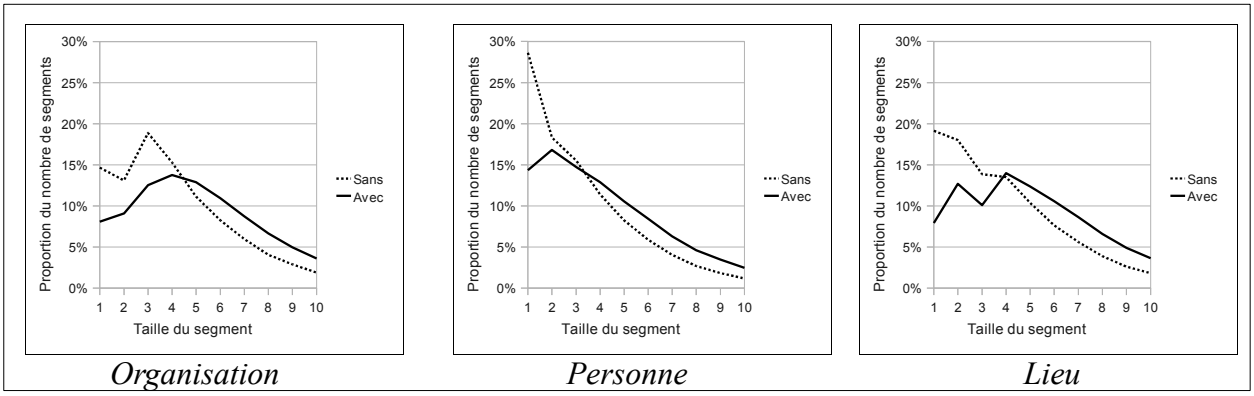


FIG 9.3 – Nombre de Segments contenant une Organisation (à gauche), une Personne (au milieu) et un Lieu (à droite) selon la taille, avant/après application de règles de segments

Le nombre de segments de taille 1 est réduit de moitié pour ce qui concerne les Personnes, de 60% pour les Lieux (plus grand écart observé) et de 45% pour les Organisations. Nous nous rapprochons donc de notre objectif : les EN sont à présent plus régulièrement associées à des segments de plus grande taille, ou contenues dans des nœuds dont ils dépendent (comme les parenthétiques par exemple). On peut supposer que les nouveaux segments auxquels les EN appartiennent contiennent des verbes avec lesquels elles sont en relation. Afin de préciser cette hypothèse, nous pouvons nous intéresser à la différence de distribution des positions des entités-R au sein des segments.

9.1.2. Distribution des entités-R en fonction de la position

Nous savons que les verbes sont faiblement présents dans des segments simples (cf. *infra* figure 9.2). On peut donc supposer que les éléments qui les accompagnent sont des arguments syntaxiques potentiels. Cette supposition nécessite de connaître la position des éléments vis-à-vis du verbe dans un segment. Or, dans la représentation initiale des segments (avant l'application de règles inter-segment), les verbes ont la particularité de se situer majoritairement en début de segment. Comme on peut le constater sur la figure (9.4), 70% des occurrences de verbes se situent en première ou en seconde position, cette dernière étant la plus probable.

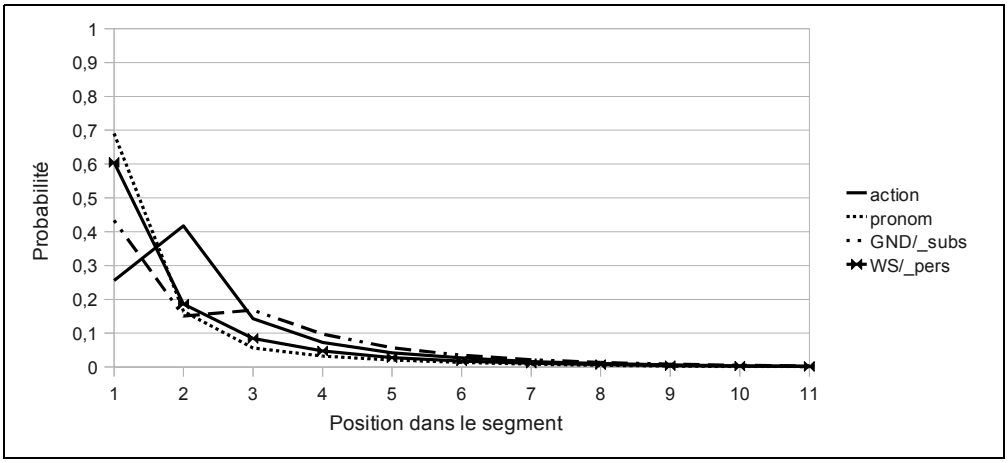


FIG 9.4 – Probabilité de position dans un segment pour les entités-R _action, _pronom GND/_subs et _pers, avant application de règles de segments

À l'inverse, les pronoms et les personnes sont fortement initiateurs de segments (0,7 et 0,6 respectivement). Ces probabilités dépendent bien entendu de la taille du segment : il n'y aurait aucun intérêt à parler de position pour un segment simple (de taille 1). Comme vu précédemment (cf. *infra* p201), les EN figurent fréquemment dans des segments simples, notamment les personnes (0,26), alors que les pronoms sont faiblement représentés (0,7).

Si l'on compare cette distribution avec celle que l'on obtient après application des règles inter-segments (figure 9.5), on observe les principaux changements suivants : la proportion de verbes apparaissant en position initiale est réduite pour moitié (0,14 contre 0,26, plus grand écart observé) alors que la position seconde n'est que légèrement augmentée (0,45 contre 0,41). Les pronoms n'ont pas changé de profil. Les entités-R de type Personne sont légèrement moins présentes en position initiale (15% de moins environ) tandis que les groupes nominaux dont la tête est un substantif sont moins fréquemment observés en position 2.

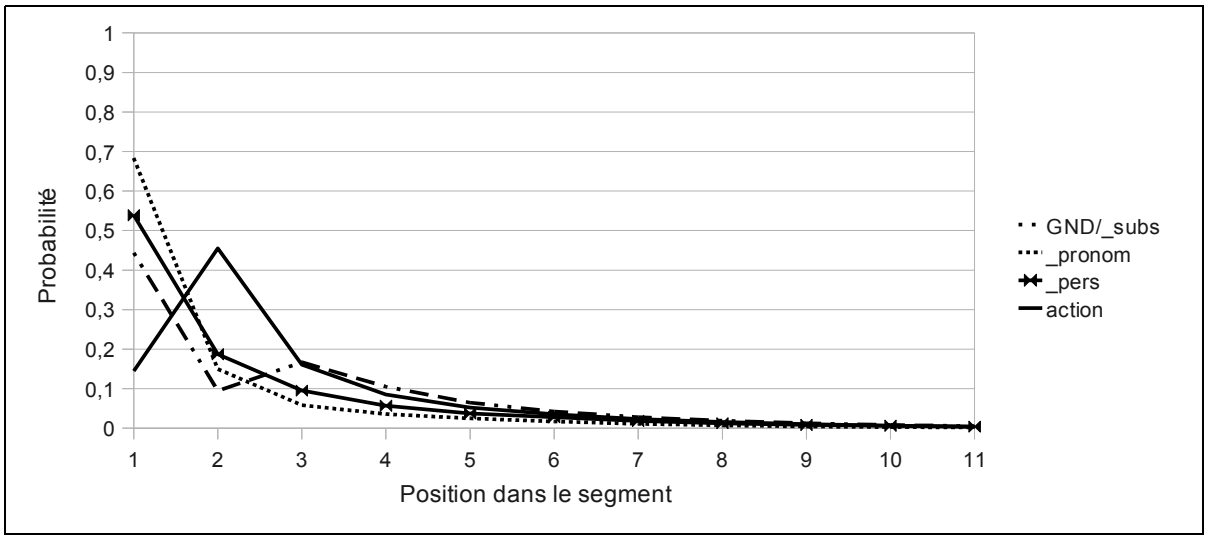


FIG 9.5 – Probabilité de position dans un segment pour les entités-R _action, _pronom GND/_subs et _pers, après application de règles de segments

La question légitime qui découle de ces observations est de savoir quelles entités-R occupent la position initiale de segments à la place du verbe et si ces entités-R sont véritablement liées sémantiquement ou syntaxiquement au verbe.

D. Bourigault, développeur de l'analyseur syntaxique ayant obtenu les meilleurs scores de précision (0,8) et de f-mesure (0,66, pour un rappel de 0,58) sur la détection de relations syntaxiques dans le corpus LeMonde lors de la campagne Easy[47], nous rappelle que :

> « Avec la relation Sujet, l'élaboration des algorithmes de reconnaissance est rendue complexe du fait de la variété des configurations que l'on peut rencontrer entre un verbe et son sujet, en particulier à cause de l'interposition possible de subordonnées relatives et de séquences incises entre virgules. »[Bourigault, 2007 : 102–103]

Le paradoxe est que c'est une des relations pour lesquelles les systèmes sont réputés obtenir en général de bonnes performances (0,92 d'après le site de la campagne, 0,83 d'après [Laurent et al., 2009]). Il faut également tenir compte du fait que c'est la relation la plus fréquente. Nous présentons dans la section suivante une grammaire qui permette la détection de cette relation. Elle nous permettra d'évaluer à la fois l'importance de chaque type d'entité-R entrant dans une telle relation, et l'apport d'un grammaire inter-segment.

47 http://www.technolangue.net/article198.html

9.2. Une grammaire syntaxique par triplet

9.2.1. Principes

La façon la plus simple de concevoir une grammaire syntaxique est d'extraire des n-grammes de catégories morpho-syntaxiques, dans notre cas des entités-R. En sélectionnant les bigrammes dont le premier est un pronom (ou un groupe nominal, etc.) et le second est un verbe, on peut par exemple espérer capter quelques relations. Deux problèmes se posent :

- la relation sujet n'est pas toujours ainsi ordonnée. Nous avons par exemple décrit les cas d'inversion de sujet pour les verbes citationnels. Leur traitement suppose de les distinguer des objets directs.
- le nombre de mots séparant le sujet de son verbe est inconnu. Travailler sur des chunks limite grandement le problème (auxiliaires des verbes et adjectifs des noms peuvent être ignorés), mais s'avère insuffisant. L'argument sujet peut être séparé par des adverbes, ou encore des appositions, des parenthétiques, etc., et ce, dans les deux directions (à gauche et à droite).

Pour éviter la profusion de patrons, règles ou automates, il semble nécessaire d'introduire une récursivité dans le parcours des nœuds de l'arbre d'une phrase qu'on pourra limiter aux frontières de phrase ou de proposition (dans notre cas des segments). Les règles ne seront alors que partiellement dépendantes de la position vis-à-vis du nœud-cible (le verbe) puisque seule la direction est connue. L'algorithme devra donc évaluer les contraintes d'une règle sur chaque nœud dans une direction donnée. Ces contraintes seront dans notre cas la nature du chunk (un GNP a très peu de chances d'être sujet à moins d'une erreur), la nature d'entité-R (un adjectif, idem) et la nature du mot (un lexique de pronoms sujets pour distinguer *il* de *lui* par exemple). Comme un segment initié par la conjonction *que* peut être une complétive (segment en position objet), nous avons enfin intégré des contraintes sur les segments, ce qui résulte en quatre types de règles. À gauche du verbe :

1. Les règles Mot : les contraintes s'établissent sur les formes des feuilles et correspondent strictement à des pronoms sujet (*je*, *tu*, *il*, etc.) des pronoms démonstratifs (*ce*, *ceci*, etc.) ou d'autres formes comme *aucun*, *chacune*. Un « lookahead » (recherche contextuelle) est également utilisé pour désambiguiser les pronoms *nous* et vous (*nous nous demandons*).
2. Les règles Chunk : les chunk de sous-type Groupe Nominal
3. Les règles Entité-R : comme certaines entités-R ne sont pas incluses dans des chunks, nous spécifions une liste possible, notamment les EN.
4. Les règles Segment : si le segment qui précède a pour frontière finale un guillemet ou si le segment dont fait partie le nœud-cible a pour frontière initiale une conjonction *que* ou *qui* (cas des relatives).

Pour le parcours à droite, les règles sont également de quatre types :

1. Les règles Chunk : les groupes nominaux, les groupes verbaux prépositionnels, les groupes nominaux prépositionnels
2. Les règles Mot : liste de verbes à l'infinitif, et de pronoms sujets inversés (*-il*, *-elle*, etc.)
3. Les règles Entité-R : les EN essentiellement

4. Les règles Segment : si le segment qui suit a pour frontière initiale un guillemet ou une conjonction *que* (cas des complétives).

En termes de représentation, les travaux en grammaire générative dans le cadre de la théorie X-bar [Chomsky, 1970] proposent de représenter les dépendances syntaxiques en terme d'antisymétrie. Par exemple, le sujet est considéré comme le spécifieur du groupe verbal de plus haut niveau V'' ou $\overline{\overline{V}}$, et l'objet, comme le complément du niveau intermédiaire V' ou $\overline{V}$. Ces concepts permettent de structurer l'arbre syntaxique et d'en déduire notamment le sujet et le complément en fonction de leur position hiérarchique. Tous les groupes peuvent a priori être représentés de cette manière, comme illustré dans une représentation très simplifiée en figure (9.6).

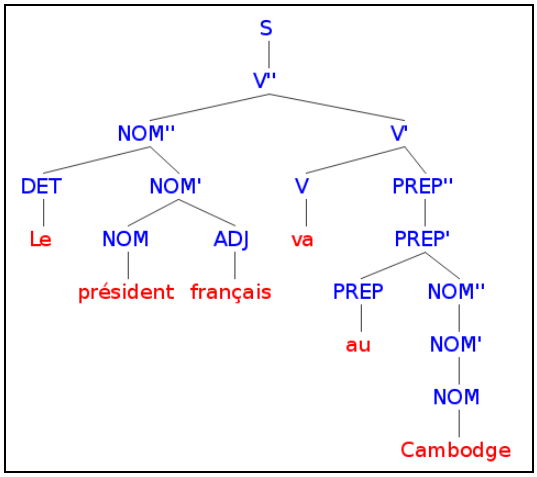


FIG 9.6 – Représentation X-bar

Comme on le voit sur la figure, la tête verbale V est sœur du groupe prépositionnel P'' qui est son complément sous V'. En revanche le sujet est le groupe nominal NOM'' (décomposé de la même manière) qui est sœur du groupe verbal de niveau intermédiaire V' sous V'' (correspondant peu ou prou à une proposition dans une phrase S). Cette représentation possède des attraits, elle est par ailleurs contestée (problèmes liés à l'antisymétrie par exemple) et a fait l'objet d'un développement particulièrement approfondi. Son avantage est de pouvoir structurer hiérarchiquement (indifféremment de la position du sujet par exemple) une proposition. Le problème est qu'elle suppose un traitement complet de la phrase, soit de pouvoir résoudre les dépendances des mots intermédiaires entre le verbe et son sujet, alors que nous cherchons uniquement à détecter les sujets. Elle constitue néanmoins une alternative intéressante à la modification unique des étiquettes ou attributs des nœuds (par exemple ajouter au nœud identifié la propriété Sujet, comme *GN_suj*), en ne modifiant pas la structure de l'arbre.

L'action des règles de la grammaire syntaxique consiste simplement à modifier les attributs du nœud validant la contrainte en y ajoutant le type de la relation (*Suj*, *Obj*, *Comp* pour les autres) et en attribuant l'identifiant de ce nœud comme valeur à l'attribut de la relation au nœud-cible (le verbe). Cette dernière caractéristique peut notamment être exploitée pour désambiguïser des cas d'inversion du sujet. La grammaire est limitée à la détection d'un argument à gauche et d'un argument à droite. On peut donc visualiser les extractions sous forme de triplet <*SUJET, VERBE, COMPLÉMENT*>. Nous décrivons dans la partie suivante les résultats obtenus pour les deux représentations du texte, avant et après l'application de règles inter-segment.

9.2.2. Triplets syntaxiques

Pour chaque occurrence verbale, nous obtenons 4 types de triplets syntaxiques possibles en fonction de la présence/absence de nœuds sélectionnés par la grammaire à gauche et à droite : *<X,V,Y>* (un élément a été trouvé à gauche et à droite), *<X,V,Ø>* (aucun élément n'a été trouvé à droite), *<Ø,V,Y>* (aucun élément n'a été trouvé à gauche) et *<Ø,V,Ø>* (aucun élément n'a été trouvé). Le tableau (9.1) illustre les différences d'extraction avec et sans application de la grammaire de segments (nous ne prenons pas en compte les relatives, ni les segments dans ce tableau) sur le corpus de développement (un peu plus d'un million d'occurrences de verbes). Nous y faisons figurer ce que nous appellerons le taux d'augmentation (ou de réduction).

Patron	Sans	Avec	Taux d'augmentation
XVY	41,8%	52,0%	24,19%
ØVØ	10,2%	6,7%	-34,32%
XVØ	18,1%	18,8%	3,77%
ØVY	29,8%	22,5%	-24,44%

Tableau 9.1 – Triplets syntaxiques obtenus avec/sans application de règles inter-segments

On observe que 42% des verbes ont été liés avec des éléments à gauche et à droite, alors que 10% n'ont été associés avec aucun élément. Lorsque l'on applique les règles inter-segment, on obtient près de 25% de plus pour les triplets *<X,V,Y>* et on réduit de près de 35% les séquences *<Ø,V,Ø>*.

Si l'on s'intéresse en détail à la nature des chunks et des entités-R qui entrent dans ces structures, on constate que l'on peut trouver des structures bien formées pour chacune. Le tableau (9.8) recense les triplets les plus fréquemment observés illustrés par des exemples (tous issus du corpus). Pour les cas où un des arguments est absent, nous avons utilisé des exemples d'impératifs, bien que l'on sache qu'ils sont moins fréquemment employés dans le corpus et qu'il s'agit généralement d'erreurs (d'étiquetage ou de détection par exemple), qu'ils sont assez souvent suivis d'une complétive ou encore précédés d'un pronom relatif.

Patron			Probabilité		Taux	Exemple		
X	V	Y	Sans	Avec		X	V	Y
_pers	V	GN/_subs	0,00696	0,01012	45,3%	*Manuel Marin*	*avouait*	*sa perplexité*
GN/_subs	V	GVPde/action	0,00428	0,00580	35,7%	*le gouvernement*	*n' essaiera pas*	*de le détourner*
GN/_subs	V	GN/_subs	0,03209	0,04344	35,4%	*les candidats*	*avancent*	*leurs solutions*
GN/_subs	V	GNPpar/subs	0,00560	0,00735	31,3%	*les divers bilans officiels*	*sont démentis*	*par les témoignages*
GN/_subs	V	GNPà/subs	0,01023	0,01338	30,8%	*ces mammifères*	*n'appartiennent pas*	*à la même famille*
GN/_subs	V	GNPdans/subs	0,00404	0,00521	28,9%	*les manifestants*	*ont défilé*	*dans la ville*
GN/_subs	V	GNPde/subs	0,01828	0,02356	28,9%	*le gaz naturel*	*a augmenté*	*de 55 pourcent*
_pers	V	Ø	0,00776	0,00894	15,2%	*Judith Butler*	*estime*	*Ø*
_pronom	V	_pronom	0,00605	0,00673	11,3%	*on*	*a vu*	*ça*
GN/_subs	V	Ø	0,05719	0,06324	10,6%	*l'abbatage*	*se poursuit*	*Ø*
Ø	V	_pers	0,01154	0,01215	5,3%	*Ø*	*assène*	*Françoise Grossetête*
_pronom	V	GVPde/action	0,00727	0,00756	4,1%	*j'*	*arrête*	*de fumer*
_pronom	V	GNPde/subs	0,01959	0,01998	2,0%	*ils*	*nécessitent*	*de la mémoire vive*
_pronom	V	GN/_subs	0,04384	0,04468	1,9%	*chacun*	*avait*	*sa zone*
_pronom	V	GNPà/subs	0,00900	0,00917	1,8%	*elle*	*répondait*	*à plusieurs objectifs*
_pronom	V	_action	0,01054	0,01069	1,5%	*il*	*fallait*	*oser*
Ø	V	_pronom	0,01651	0,01649	-0,1%	*Ø*	*oublions*	*ça*
_pronom	V	Ø	0,07146	0,06703	-6,2%	*il*	*arrive*	*Ø*
Ø	V	_action	0,00565	0,00517	-8,5%	*Ø*	*osons*	*oser*
Ø	V	GN/_subs	0,07770	0,05793	-25,4%	*Ø*	*arrive*	*un nouveau convive*
Ø	V	GNPde/subs	0,03340	0,02441	-26,9%	*Ø*	*saupoudrez*	*de persil haché*
Ø	V	GVPde/action	0,00790	0,00562	-28,8%	*Ø*	*feignons*	*de le croire*
Ø	V	GNPdans/subs	0,00712	0,00479	-32,7%	*Ø*	*marchez*	*dans les rues*
Ø	V	GNPà/subs	0,01694	0,01123	-33,7%	*Ø*	*revenons*	*à nos épigraphes*
Ø	V	Ø	0,10236	0,06723	-34,3%	*Ø*	*recommencez*	*Ø*
Ø	V	GNPpar/subs	0,00767	0,00460	-40,0%	*Ø*	*commençons*	*par la seconde question*

Tableau 9.2 – Triplets obtenus illustré d'exemples avec/sans application de règles inter-segments

Le tableau (9.2) fait également figurer les probabilités d'observer ces triplets parmi l'ensemble des possibles (la totalité des occurrences des verbes). Le taux (d'augmentation ou de réduction) entre les triplets extraits avec et sans application de la grammaire de segment est également indiqué. Les triplets sont classés selon ce taux. On constate que les triplets complets (*<X,V,Y>*) sont constamment plus fréquents après application de cette grammaire ; par exemple on obtient 45% de plus de triplets *<_pers, V, GN/_subs>*. Cette augmentation est associée à une réduction des structures où un argument du triplet est manquant ; par exemple, le triplet *<Ø,V,Ø>* est réduit de 35%, conformément au tableau (9.1).
Pour plus de précisions, nous avons fait figurer les éléments les plus fréquemment rencontrés à gauche (tableau 9.3) et à droite (tableau 9.4) du verbe en indiquant les taux comme précédemment.

X	Sans	Avec	Taux
GN/_pers_fonct	0,0134	0,0167	24,9%
_pers	0,0483	0,0565	16,9%
_np	0,0180	0,0206	14,5%
GN/_np	0,0128	0,0146	14,0%
GN/_pers_act	0,0104	0,0117	12,5%
_org	0,0149	0,0166	11,5%
GN/_pers	0,0137	0,0149	8,6%
GN/_org	0,0222	0,0241	8,6%
GN/_subs	0,3170	0,3390	6,9%
_pronom	0,3807	0,3221	-15,4%

Tableau 9.3 – Détail des entités-R obtenues à gauche (X) avec/sans application de règle de segment

Y	Sans	Avec	Taux
_pronom	0,0408	0,0435	6,8%
_pers	0,0200	0,0214	6,6%
_action	0,0315	0,0329	4,6%
GVPde/action	0,0344	0,0353	2,6%
GVPà/action	0,0212	0,0214	0,7%
GN/_subs	0,2670	0,2689	0,7%
GNPde/subs	0,1197	0,1183	-1,1%
GNPdans/subs	0,0258	0,0253	-1,9%
GNPà/subs	0,0616	0,0603	-2,1%
GNPpar/subs	0,0253	0,0242	-4,1%

Tableau 9.4 – Détail des entités-R obtenues à droite (Y) avec/sans application de règle de segment

Lorsque la grammaire syntaxique trouve un élément à gauche, il s'agit à près de 40% d'un pronom et à 30% d'un groupe nominal dont la tête est un substantif. Les EN sont fortement représentées dans cette position, ainsi que les groupes nominaux dont la tête est une fonction (*le président*) ou une activité (*l'acteur*). On observe que ces EN sont plus fortement représentées après application des segments : on trouve par exemple comparativement plus de personnes à gauche (17%).
En revanche la proportion de pronoms est réduite. Cette réduction s'explique par un bug de la grammaire de segments. Nous avons permis de rattacher des pronoms en insertion sans contrôler la forme (*il*, *elle*, par opposition à *lui*, *elle*) ainsi que les expressions temporelles : cette réduction est due à leur adjacence et la règle appliquée en priorité est celle du rattachement de pronom au temps plutôt que l'inverse ; étant donné le temps de traitement, nous n'avons pas pu inclure cette rectification dans les résultats qui suivent.

Ces 10 éléments représentent près de 85% des éléments identifiés par la grammaire (l'application de la grammaire joue peu de ce point de vue). Pour ce qui concerne la position droite (Y), les 10 groupes les plus fréquents représentent 65% de la totalité. On observe moins de différences provoquées par l'application des règles inter-segments. Les groupes les plus fortement représentés sont les groupes nominaux dont l'entité-R tête est un substantif, les groupes prépositionnels nominaux en *de* et en *à*. Il ne reste à présent plus qu'à évaluer la pertinence de ces triplets du point de vue des relations syntaxiques.

9.3. Évaluation de la détection de sujet à longue distance

Pour évaluer l'apport de la grammaire inter-segment dans la détection de sujet à longue distance, nous n'avons bien évidemment pas pu consulter tous les triplets extraits. Nous nous sommes concentré sur un verbe particulier, *annoncer*, choisi pour sa fréquence et la variabilité de ses structures. Nous décrirons dans un premier temps le comportement de ce verbe, puis les triplets obtenus et enfin les résultats de l'évaluation

9.3.1. Caractérisation syntaxico-sémantique du verbe *annoncer*

Le verbe *annoncer* est suffisamment fréquent (environ 6000 occurrences de ses formes dans la totalité du corpus), il se combine avec une variété de types sémantiques et possède une large gamme de structures. Ce verbe peut être employé sur le mode pronominal (220), actif (221), passif (222) et causatif (223).

(220)le congrès de novembre 2004 *s'annonce* comme le rendez-vous de l'année, pour l'UMP.
(221)plusieurs sociétés d' électronique *doivent annoncer* prochainement leurs résultats pour le dernier trimestre 2003
(222)sa mort *est annoncée* seulement lors de la fête organisée après le spectacle
(223)il *fait annoncer* la fin de la trêve pour le lendemain mardi 22 août

On observe que les formes pronominale et passive ont pour sujet des événements (*congrès*, *mort*) et que les formes actives et causatives sont plutôt des agents, bien qu'on ne puisse réduire cette catégorie aux Humains. Ces quatre exemples types entrent dans les structures syntaxiques majeures suivantes :

- les formes actives
  - [GN V que] : structure dans laquelle l'objet est une complétive

    (224)Matthew Cooper a annoncé que son journal fera appel

  - [Citation V GN] : structure citationnelle où le sujet est inversé

    (225)l'Italie et l'Allemagne préparent une initiative commune pour lutter contre l'immigration clandestine, a annoncé, jeudi 12 août, le ministre italien de l'intérieur

  - [GN V GN] : structure dans laquelle l'objet est un groupe nominal

    (226)une organisation grecque de gauche a annoncé la tenue d'une manifestation, vendredi 27 août

  - [GN V Infinitif] : structure dans laquelle l'objet est un infinitif. On peut considérer que le verbe *annoncer* joue ici un rôle mineur, de l'ordre de l'auxiliaire ou du modal.

    (227)Renaud Dutreil, le ministre de la fonction publique, a annoncé vouloir diviser par deux le coût des logiciels de l'Etat en ayant recours au système libre

- les formes pronominales
  - [GN V Adjectif] : structure qui s'apparente à une relation attributive (remplacer le verbe de 226 par *sont*)

(228)les futurs rapports financiers s'annoncent excellents

  - [Date V Manière] : structure où le sujet est un événement ou une date ; la manière peut être représentée sous divers formes (par un adverbe ou un SP notamment)

(229)2004 ne s'annonce pas sous de meilleurs auspices pour l'emploi et les salaires

Les formes causatives et passives sont majoritairement des variantes des formes actives que nous n'avons pas fait figurer ici. La forme passive a la particularité de pouvoir être employée intransitivement. Il faut également recenser des formules idiomatiques telles que *annoncer la couleur*, et le fait qu'on rencontre assez souvent des inversions du sujet. L'analyse syntaxique de ces structures semble au premier abord complexe.

Si on s'intéresse au plan sémantique, les sujets des verbes sont régulièrement des EN, mais leur nature est variable : organisations, événements, personnes, supports médiatiques, phénomènes, etc. Ces EN peuvent être employées métonymiquement, ce qui peut compliquer des tentatives de contrôle sémantique. On trouve en revanche rarement des lieux. Les objets regroupent des catégories encore plus diverses, puisque l'on peut annoncer toutes sortes de choses : des noms, des événements, des intentions, des projets, des bornes temporelles, etc. Il ne semble donc pas y avoir de corrélation entre propriétés syntaxiques et sémantiques.

Néanmoins, si on se concentre sur les formes actives, on peut réduire le champ des possibles : lorsque le verbe est employé dans une structure [GN V GN], les objets sont rarement des noms d'agent et plus régulièrement des nominalisations. Il semble donc que les EN puissent être employées pour détecter les sujets, qu'ils apparaissent à gauche ou à droite du verbe (dans des cas d'inversion du sujet). C'est ce que nous étudierons en évaluant la qualité de détection de sujet par catégorie.

9.3.2. Triplets syntaxiques

En excluant les formes réflexives, les groupes verbaux prépositionnels, ainsi que les occurrences se trouvant dans des relatives (la grammaire n'est pas paramétrée pour rechercher les antécédents des pronoms relatifs), nous obtenons 5309 occurrences (87% de la totalité). Nous avons appliqué les règles de grammaire syntaxique sur deux corpus, avant application de la grammaire de segments (modèle M1) et après (modèle M2). Les types de triplets retournés par la grammaire avant l'application des règles inter-segment sont indiqués tableau (9.5).

Triplet	Occurrences totales			Hapax		Absent	
	Avant	Après	Taux	Avant	Après	Avant	Après
XVY	29,6%	45,9%	54,8%	6,0%	9,1%	4,2%	0,1%
ØVØ	18,2%	8,7%	-52,4%	0,0%	0,0%	0,0%	0,0%
ØVY	19,4%	20,9%	7,4%	1,1%	1,1%	0,3%	0,2%
XVØ	32,7%	24,5%	-24,9%	0,2%	0,3%	0,0%	0,0%

Tableau 9.5 – Triplets syntaxiques obtenus avec/sans application de règles inter-segments pour le verbe annoncer

On constate que le modèle M2 obtient plus de 50% de plus de triplets complets (45,9% contre 29,6%) et qu'il réduit d'autant les triplets <*Ø,V,Ø*>. Si l'on compare ces taux avec les données obtenues sur la totalité des verbes (tableau 9.1 ; cf. *infra* p222), on observe que ce verbe est beaucoup plus influencé par l'application de règles inter-segment. Nous avons également fait figurer la proportion de triplets hapax (n'apparaissant qu'une fois pour un modèle donné) que l'on retrouve essentiellement pour les triplets de type <*X,V,Y*>, la proportion d'hapax communs étant de 5%. La majorité des hapax obtenus par le modèle M2 qui ne sont pas communs sont absents du modèle M1. Nous avons indiqué dans le tableau (9.6) les 10 triplets <*XVY*> communs les plus fréquents dans le modèle M2 (ce sont également les triplets les plus fréquents de M1, hormis un seul dont le taux est de +7% pour le modèle M2).

Patron		Fréquence		Probabilité		Taux
X	Y	M1	M2	M1	M2	
GN/_subs	*GN/_subs*	160	233	0,0301	0,0439	46%
_pronom	*GN/_subs*	155	131	0,0292	0,0247	-15%
_pers	*GN/_subs*	67	119	0,0126	0,0224	78%
GN/_Org	*GN/_subs*	39	109	0,0073	0,0205	179%
GN/_pers_fonct	*GN/_subs*	52	92	0,0098	0,0173	77%
_org	*GN/_subs*	46	88	0,0087	0,0166	91%
GN/_org	*GN/_subs*	27	78	0,0051	0,0147	189%
GN/_subs	*GNPde/subs*	54	65	0,0102	0,0122	20%
GN/_pers	*GN/_subs*	46	55	0,0087	0,0104	20%
_np	*GN/_subs*	24	53	0,0045	0,0100	121%

Tableau 9.6 – Détail des Triplets syntaxiques obtenus par les modèles M1 et M2 pour <X,V,Y>

Ce tableau indique que ces triplets sont plus nombreux pour le modèle M2, résultant en un taux d'augmentation pouvant atteindre les 189%, pour le triplet <*GN/_org, annoncer, GN/_subs*>, dont (230) est un exemple.

(230)le groupe Publicis a annoncé, jeudi 8 janvier, la création de Publicis Events Worldwide

Indiquons que les meilleurs taux (qui ne figurent pas dans ce tableau) peuvent atteindre 600% (ce qui excède largement la moyenne), comme pour le triplet <*GN/_subs, annoncer, GN/_org*>, de fréquence 7 pour le modèle M2 et de 1 pour le modèle M1. On remarque une fois de plus que les triplets dont X est un pronom sont moins fréquents pour le modèle M2. Observons que la position Y est quasi-systématiquement occupée par des GN dont la tête est un substantif et que la position X est fortement représentée par des Npr (environ la moitié) : ceci peut s'expliquer par le fait qu'un grand nombre de règles de la grammaire de segments s'appliquent aux segments dont le nœud final est une EN.

Pour être complet, nous reportons les données obtenues pour les deux autres types de triplets (tableaux 9.7 et 9.8), toujours sur les mêmes critères d'ordonnancement.

Patron		Fréquence		Probabilité		Taux
X	Y	M1	M2	M1	M2	
Ø	*GN/_subs*	406	427	0,0765	0,0804	5%
Ø	*GN/_pers_fonct*	61	79	0,0115	0,0149	30%
Ø	*_pers*	51	53	0,0096	0,0100	4%
Ø	*GNPde/subs*	50	49	0,0094	0,0092	-2%
Ø	*GNPà/subs*	26	25	0,0049	0,0047	-4%
Ø	*GN/_np*	24	25	0,0045	0,0047	4%
Ø	*GN/_Org*	22	25	0,0041	0,0047	14%
Ø	*_pronom*	23	21	0,0043	0,0040	-9%
Ø	*GNPpar/subs*	26	20	0,0049	0,0038	-23%
Ø	*GN/_pers*	19	19	0,0036	0,0036	0%
Ø	*GN/_org*	12	18	0,0023	0,0034	50%

Tableau 9.7 – Détail des Triplets syntaxiques obtenus par les modèles M1 et M2 pour <Ø,V,Y>

Patron		Fréquence		Probabilité		Taux
X	Y	M1	M2	M1	M2	
GN/_subs	Ø	423	384	0,0797	0,0723	-9%
_pronom	Ø	243	181	0,0458	0,0341	-26%
_pers	Ø	126	104	0,0237	0,0196	-17%
GN/_Org	Ø	118	36	0,0222	0,0068	-69%
GN/_org	Ø	110	62	0,0207	0,0117	-44%
_org	Ø	103	61	0,0194	0,0115	-41%
GN/_pers_fonct	Ø	86	66	0,0162	0,0124	-23%
GN/_pers	Ø	69	56	0,0130	0,0105	-19%
_pers_bof	Ø	58	52	0,0109	0,0098	-10%
GN/_org_prob	Ø	51	22	0,0096	0,0041	-57%
_np	Ø	35	20	0,0066	0,0038	-43%
GN/_np	Ø	31	29	0,0058	0,0055	-6%

Tableau 9.8 – Détail des Triplets syntaxiques obtenus par les modèles M1 et M2 pour <X,V,Ø>

La réduction que l'on observe pour 4 triplets dans le tableau (9.7) peut s'expliquer par la raison suivante : les triplets *<Ø,V,Y>* ont disparu parce qu'un élément X (juste ou erroné) a été identifié par le modèle M2. Cette redistribution ne signifie cependant pas que les entités-R identifiées en position Y pour ce type de triplets ne soient pas de bons candidats sujet. L'augmentation, en revanche, signifie que là où le modèle M1 n'identifiait aucun élément X ou Y, l'application de la grammaire de segmentation a permis d'identifier des éléments supplémentaires en position Y. On constate une fois de plus que la moitié des triplets fréquents contiennent des EN, pour lesquelles le modèle M2 augmente le nombre.

Pour ce qui concerne les triplets les plus fréquents de type *<X,V,Ø>* (tableau 9.8), tous sont associés à une diminution, la plus forte étant enregistrée pour le triplet *<GN/_Org, annonce, Ø>*. Les mêmes remarques que nous venons de faire pour les triplets *<Ø,V,Y>* s'appliquent. Ajoutons pour finir que l'absence d'éléments à droite et à gauche peut s'expliquer par les constructions spécifiques dans lesquelles entre ce verbe : il peut être suivi d'une complétive, auquel cas il s'agit d'un segment initié en *que*, ou, être précédé, voire suivi, d'une citation, auquel cas ces segments sont inclus entre guillemets.

Combien d'éléments supplémentaires le modèle M2 a-t-il permis d'identifier pour chaque catégorie de triplet (X ou Y) ? Le modèle M2 a permis d'extraire 516 nouveaux éléments en position X et 894 en position Y. Comme l'illustrent les tableaux (9.9) et (9.10), ordonnés en fonction du nombre de nouveaux candidats supérieur à 5, pour chaque position, les groupes nominaux dont la tête est un substantif représentent une grande part de ces nouveaux candidats (47% en X et Y).

Position X	Fréquence		Probabilité		Taux	Nouveaux Candidats
	M1	M2	M1	M2		
GN/_subs	967	1109	0,1821	0,2089	15%	142
_pers	221	287	0,0416	0,0541	30%	66
GN/_pers_fonct	157	195	0,0296	0,0367	24%	38
_np	68	94	0,0128	0,0177	38%	26
_org	179	203	0,0337	0,0382	13%	24
GN/_org	159	178	0,0299	0,0335	12%	19
GN/_Org	174	192	0,0328	0,0362	10%	18
_loc	16	34	0,0030	0,0064	113%	18
GN/_org_prob	62	76	0,0117	0,0143	23%	14
GN/_pers_act	29	43	0,0055	0,0081	48%	14
GN/_loc	10	22	0,0019	0,0041	120%	12
_pers_bof	76	87	0,0143	0,0164	14%	11
GN/_np	53	64	0,0100	0,0121	21%	11
GN/_val_subs	23	31	0,0043	0,0058	35%	8
GN/_Organisation	59	65	0,0111	0,0122	10%	6
GN/_fp	42	48	0,0079	0,0090	14%	6

Tableau 9.9 – Détail des entités-R obtenues à gauche (X) par les modèles M1 et M2

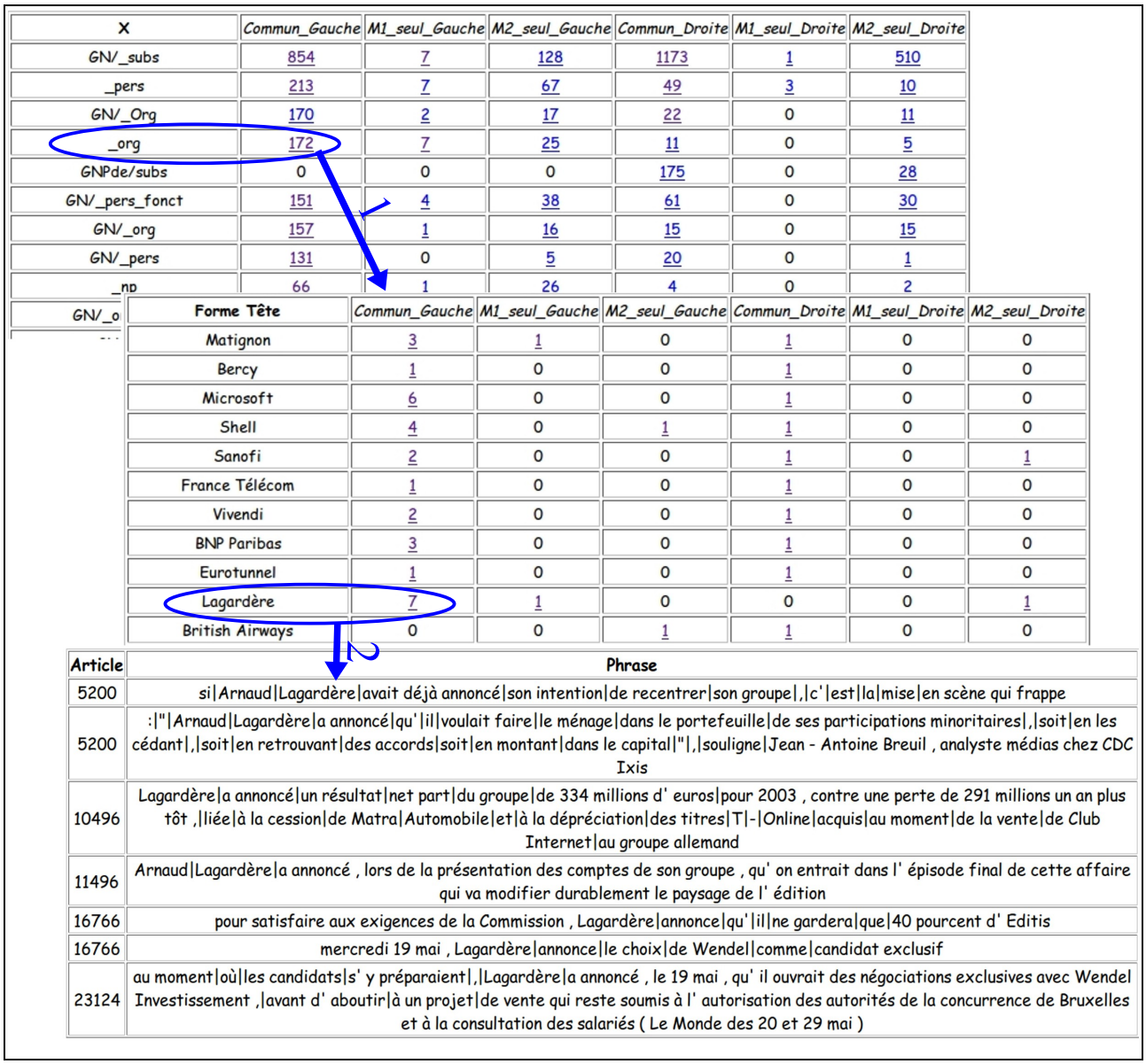
Position Y	Fréquence		Probabilité		Taux	Nouveaux Candidats
	M1	M2	M1	M2		
GN/_subs	1183	1699	0,2228	0,3200	44%	516
GN/_pers_fonct	61	92	0,0115	0,0173	51%	31
GNPde/subs	178	207	0,0335	0,0390	16%	29
GN/_score	12	30	0,0023	0,0057	150%	18
GNPà/subs	76	92	0,0143	0,0173	21%	16
GN/_Qnombre	24	40	0,0045	0,0075	67%	16
GNPdans/subs	50	65	0,0094	0,0122	30%	15
_action	32	47	0,0060	0,0089	47%	15
GN/_org	15	30	0,0028	0,0057	100%	15
GNPà/loc	22	36	0,0041	0,0068	64%	14
GN/_type_loi	17	31	0,0032	0,0058	82%	14
GN/_Org	22	33	0,0041	0,0062	50%	11
GN/_org_prob	7	16	0,0013	0,0030	129%	9
_pers	52	59	0,0098	0,0111	13%	7
GNPpar/subs	51	58	0,0096	0,0109	14%	7
GN/_acro_div	13	19	0,0024	0,0036	46%	6

Tableau 9.10 – Détail des entités-R obtenues à droite (Y) par les modèles M1 et M2

Il y a en tout 75 types d'éléments possibles en position X et 223 en position Y. Cette différence s'explique par le fait que la grammaire syntaxique autorise un plus grande diversité d'éléments à droite (elle inclut les GVP et les GNP qui varient selon les nature de l'entité-R tête et de la préposition). 47 éléments en position X et 85 éléments en position Y ont au moins un nouveau candidat, soit respectivement 62% et 38%. Le reste est invariant, hormis les pronoms pour la position X (85 occurrences en moins pour un taux de -16%) et un hapax en position Y. On constate avec intérêt l'augmentation des entités-R *_score* et *_Qnombre* qui correspondent respectivement à des valeurs (*1,5 pourcent*, etc.) et à des noms de calcul (*augmentation*, *hausse*, *baisse*, etc.), mais nous n'évaluons pas la détection des compléments objets.

9.3.3. Résultats

Pour réaliser l'évaluation, nous avons conçu une interface de visualisation web qui permet de consulter, à partir de chaque verbe, les éléments trouvés à gauche et à droite en fonction de leur catégorie et type de chunk, ainsi que lorsque rien n'a été trouvé par la grammaire syntaxique (les éléments des triplets). Chaque catégorie est reliée par un lien inter-texte à une page où figurent les formes têtes observées dans ces positions, qui sont elles-mêmes liées à leurs occurrences en contexte. La figure (9.7) illustre les liens entre les différentes pages de cette interface.



X	Commun_Gauche	M1_seul_Gauche	M2_seul_Gauche	Commun_Droite	M1_seul_Droite	M2_seul_Droite
GN/_subs	854	7	128	1173	1	510
_pers	213	7	67	49	3	10
GN/_Org	170	2	17	22	0	11
_org	172	7	25	11	0	5
GNPde/subs	0	0	0	175	0	28
GN/_pers_fonct	151	4	38	61	0	30
GN/_org	157	1	16	15	0	15
GN/_pers	131	0	5	20	0	1
_np	66	1	26	4	0	2

GN/_o

Forme Tête	Commun_Gauche	M1_seul_Gauche	M2_seul_Gauche	Commun_Droite	M1_seul_Droite	M2_seul_Droite
Matignon	3	1	0	1	0	0
Bercy	1	0	0	1	0	0
Microsoft	6	0	0	1	0	0
Shell	4	0	1	1	0	0
Sanofi	2	0	0	1	0	1
France Télécom	1	0	0	1	0	0
Vivendi	2	0	0	1	0	0
BNP Paribas	3	0	0	1	0	0
Eurotunnel	1	0	0	1	0	0
Lagardère	7	1	0	0	0	1
British Airways	0	0	1	1	0	0

Article	Phrase
5200	si|Arnaud|Lagardère|avait déjà annoncé|son intention|de recentrer|son groupe|,|c'|est|la|mise|en scène qui frappe
5200	:|"|Arnaud|Lagardère|a annoncé|qu'|il|voulait faire|le ménage|dans le portefeuille|de ses participations minoritaires|,|soit|en les cédant|,|soit|en retrouvant|des accords|soit|en montant|dans le capital|"|,|souligne|Jean - Antoine Breuil , analyste médias chez CDC Ixis
10496	Lagardère|a annoncé|un résultat|net part|du groupe|de 334 millions d' euros|pour 2003 , contre une perte de 291 millions un an plus tôt ,|liée|à la cession|de Matra|Automobile|et|à la dépréciation|des titres|T|-|Online|acquis|au moment|de la vente|de Club Internet|au groupe allemand
11496	Arnaud|Lagardère|a annoncé , lors de la présentation des comptes de son groupe , qu' on entrait dans l' épisode final de cette affaire qui va modifier durablement le paysage de l' édition
16766	pour satisfaire aux exigences de la Commission , Lagardère|annonce|qu'|il|ne gardera|que|40 pourcent d' Editis
16766	mercredi 19 mai , Lagardère|annonce|le choix|de Wendel|comme|candidat exclusif
23124	au moment|où|les candidats|s' y préparaient|,|Lagardère|a annoncé , le 19 mai , qu' il ouvrait des négociations exclusives avec Wendel Investissement ,|avant d' aboutir|à un projet|de vente qui reste soumis à l' autorisation des autorités de la concurrence de Bruxelles et à la consultation des salariés (Le Monde des 20 et 29 mai)

FIG 9.7 – Exemples de pages liées dans l'interface de visualisation des triplets syntaxiques

Dans la première page, sont triées les occurrences du verbe en fonction de ses cooccurrences selon un modèle de sélection donné. En sélectionnant une catégorie et un modèle, l'utilisateur est renvoyé

vers la liste des formes apparaissant dans ce contexte. Enfin, l'utilisateur peut consulter les phrases dans lesquelles chaque forme a été identifiée. Ces phrases indiquent entre barres obliques les chunks apparaissant dans une phrase (segments à frontières fortes).

Pour évaluer les triplets, nous avons tenu compte des conventions suivantes :

- Si l'entité-R extraite fait partie du syntagme sujet, même si le chunk est mal détecté, l'exemple est validé : nous n'évaluons pas les performances du chunking, mais uniquement la capacité de la grammaire syntaxique à repérer un sujet. Précisons que si l'entité-R fait partie d'un GNP dépendant du sujet, comme X en (231), il s'agit d'une erreur ; si l'entité-R fait partie d'une apposition, il s'agit également d'une erreur (comme X en 232). Seuls sont retenus les cas où l'élément fait partie de la dénomination d'une EN comme X en (233) ou qu'il est immédiatement sous la dépendance de la tête du syntagme, comme X en (234).

  (231)[la vente] [du groupe] [X] [a été annoncée]
  (232)[le PDG de Y], [X] [a annoncé]
  (233)[Pietr] [X] a annoncé
  (234)[Le groupe] [X] [a annoncé]

- En épluchant les exemples, nous nous sommes rendu compte de certaines erreurs d'étiquetage grammatical : *annoncer* peut être employé comme nom (*une annonce*), ou participe adjectival (*les mesures annoncées*). Ces formes ont été écartées.
- Les formes participiales non finies apparaissant dans des appositions (235) et les infinitifs (236) ont été écartées, car la relation qu'ils peuvent entretenir avec un nom potentiel ne relève pas de la relation syntaxique sujet (bien qu'ils soient fréquemment sémantiquement liés). L'annotation diffère en cela de celle que nous avions établie pour le corpus de contes (cf. *infra* 3.3.1), dans laquelle nous recherchions le sujet logique plutôt que grammatical.

  (235)Annoncée comme une bombe, An Antigone, de Wanda Golonka n'a été qu'une bombinette.
  (236)C'était un grand moment de télévision, sur CNN, de voir George Bush annoncer la démission de George Tenet, directeur de la CIA depuis sept ans, puis aussitôt prendre l'avion pour Rome, sans s'attarder davantage et surtout sans répondre à une seule question.

- Enfin, pour ne pas pénaliser le modèle M2 concernant le bug des pronoms, nous avons validé les exemples où le rattachement avait été mal effectué (problème qui n'apparaît pas directement sur l'interface).

Les occurrences restantes sont au nombre de 4483 et celles qui étaient communes aux deux modèles ont été regroupées. Nous avons voulu savoir simultanément si :

- un modèle intra-segment (M1) obtenait de bonnes performances et pour quelles entités-R ces performances étaient les plus précises
- le modèle M2 contribuait à améliorer les résultats et pour quelles entités-R ses performances étaient plus précises.

Le tableau (9.11) résume les résultats des entités-R évaluées.

Mod.	Pos.	Catégorie	Err.	Bon	Ret.	Bon cum.	Ret. cum.	Préc.	Préc. cum.	Rap. cum.	F-m cum.
M1	Gauche	*GN/_subs*	42	488	530	488	530	0,921	0,921	0,109	0,195
		_pronom	28	448	476	936	1006	0,941	0,930	0,209	0,341
		_pers	0	210	210	1146	1216	1	0,942	0,256	0,402
		_org	7	164	171	1310	1387	0,959	0,944	0,292	0,446
		GN/_Org	0	169	169	1479	1556	1	0,951	0,330	0,490
		GN/_org	4	151	155	1630	1711	0,974	0,953	0,364	0,526
		_pers_fonct	1	145	146	1775	1857	0,993	0,956	0,396	0,560
		GN/_pers	0	127	127	1902	1984	1	0,959	0,424	0,588
		_pers_bof	57	18	75	1920	2059	0,240	0,932	0,428	0,587
		_np	5	57	62	1977	2121	0,919	0,932	0,441	0,599
		GN/_org_prob	0	59	59	2036	2180	1	0,934	0,454	0,611
		GN/_np	0	49	49	2085	2229	1	0,935	0,465	0,621
		GN/_fp	1	40	41	2125	2270	0,976	0,936	0,474	0,629
		GN/_pers_act	1	28	29	2153	2299	0,966	0,936	0,480	0,635
		GN/_acro_div	0	29	29	2182	2328	1	0,937	0,487	0,641
		_prod	0	24	24	2206	2352	1	0,938	0,492	0,645
		_acro_div	0	19	19	2225	2371	1	0,938	0,496	0,649
		_loc	1	11	12	2236	2383	0,917	0,938	0,499	0,651
	Droite	*_pers_fonct*	0	61	61	2297	2444	1	0,940	0,512	0,663
		_pers	1	48	49	2345	2493	0,980	0,941	0,523	0,672
		GN/_Org	0	22	22	2367	2515	1	0,941	0,528	0,676
		GN/_np	0	21	21	2388	2536	1	0,942	0,533	0,680
		GN/_pers	0	19	19	2407	2555	1	0,942	0,537	0,684
		_loc	7	10	17	2417	2572	0,588	0,940	0,539	0,685
		GN/_org	3	12	15	2429	2587	0,800	0,939	0,542	0,687
		GN/_acro_div	3	10	13	2439	2600	0,769	0,938	0,544	0,689
		_org	1	10	11	2449	2611	0,909	0,938	0,546	0,690
		GN/_pers_act	1	10	11	2459	2622	0,909	0,938	0,548	0,692
		_prod	0	8	8	2467	2630	1	0,938	0,550	0,694
		GN/_org_prob	0	6	6	2473	2636	1	0,938	0,552	0,695
		GN/_fp	0	5	5	2478	2641	1	0,938	0,553	0,696
		_np	1	2	3	2480	2644	0,667	0,938	0,553	0,696
		_pers_bof	0	3	3	2483	2647	1	0,938	0,554	0,696
		_acro_div	0	2	2	2485	2649	1	0,938	0,554	0,697
M2	Gauche	*GN/_subs*	44	62	106	2547	2755	0,585	0,925	0,568	0,704
		_pers	10	55	65	2602	2820	0,846	0,923	0,580	0,712
		_pers_fonct	2	36	38	2638	2858	0,947	0,923	0,588	0,719
		_org	4	21	25	2659	2883	0,840	0,922	0,593	0,722
		_np	7	17	24	2676	2907	0,708	0,921	0,597	0,724
		GN/_Org	1	16	17	2692	2924	0,941	0,921	0,600	0,727
		GN/_org	1	15	16	2707	2940	0,938	0,921	0,604	0,729
		GN/_pers_act	2	12	14	2719	2954	0,857	0,920	0,606	0,731
		GN/_org_prob	0	13	13	2732	2967	1	0,921	0,609	0,733
		GN/_np	0	11	11	2743	2978	1	0,921	0,612	0,735
		_pers_bof	3	5	8	2748	2986	0,625	0,920	0,613	0,736
		_loc	7	0	7	2748	2993	0	0,918	0,613	0,735
		GN/_fp	0	6	6	2754	2999	1	0,918	0,614	0,736
		GN/_pers	0	4	4	2758	3003	1	0,918	0,615	0,737
		GN/_acro_div	0	4	4	2762	3007	1	0,919	0,616	0,737
		_acro_div	1	3	4	2765	3011	0,750	0,918	0,617	0,738
		_prod	0	4	4	2769	3015	1	0,918	0,618	0,738
		_pronom	1	2	3	2771	3018	0,667	0,918	0,618	0,739
	Droite	*_pers_fonct*	3	26	29	2797	3047	0,897	0,918	0,624	0,743
		GN/_org	2	13	15	2810	3062	0,867	0,918	0,627	0,745
		GN/_Org	0	11	11	2821	3073	1	0,918	0,629	0,747
		_pers	1	8	9	2829	3082	0,889	0,918	0,631	0,748
		GN/_org_prob	1	8	9	2837	3091	0,889	0,918	0,633	0,749
		GN/_acro_div	1	5	6	2842	3097	0,833	0,918	0,634	0,750
		_org	0	5	5	2847	3102	1	0,918	0,635	0,751
		GN/_np	0	5	5	2852	3107	1	0,918	0,636	0,751
		GN/_fp	0	4	4	2856	3111	1	0,918	0,637	0,752
		GN/_pers_act	1	3	4	2859	3115	0,750	0,918	0,638	0,752
		_np	1	1	2	2860	3117	0,500	0,918	0,638	0,753
		GN/_pers	0	1	1	2861	3118	1	0,918	0,638	0,753
		_pers_bof	0	1	1	2862	3119	1	0,918	0,638	0,753
		_acro_div	1	0	1	2862	3120	0	0,917	0,638	0,753
		_loc	1	0	1	2862	3121	0	0,917	0,638	0,753
		_prod	0	0	0	2862	3121	na	0,917	0,638	0,753

Tableau 9.11 – Résultats obtenus pour l'évaluation des sujets du verbe annoncer

Le tableau (9.11) est organisé selon le type de modèle et les positions gauche ou droite sélectionnées pour l'évaluation. Il indique pour chacune de ces informations, le nombre d'erreurs (Err.), le nombre de bons candidats (Bons), le nombre d'éléments retournés (Rét.) et la précision par catégorie. Les catégories sont ordonnées en fonction du nombre d'éléments retournés. Nous avons également illustré le nombre de bons candidats cumulés (Bon cum.), le nombre d'éléments retournés (Rét. cum.), la précision cumulée (Préc. cum.), le rappel cumulé (Rap. cum.) et la F-mesure cumulée (F-m cum.).
Une grammaire qui n'identifierait que les éléments à gauche au sein d'un segment (M1 Gauche), obtiendrait 0,65 de F-mesure en ne traitant que les catégories sélectionnées, pour la moitié des occurrences totales de sujet. Dans ce modèle, la détection des groupes nominaux et des pronoms contribuerait pour 0,34 de f-mesure, les EN (nous y incluons les expressions référentielles désignées par *_pers_act*, *_pers_fonct* et *_Organisation*) pour une proportion relativement similaire (0,31). La précision de certaines catégories est excellente (1 pour l'entité-R *_pers* par exemple), mais elle peut également être médiocre : 55 des erreurs de la catégorie *_pers_bof* sont provoquées par un bug d'étiquetage. Les noms propres (*_np*) et les acronymes (*_acro_div*) contribuent également à une bonne détection, même s'ils ne sont pas sémantiquement annotés. Enfin, les lieux, sont rares mais sont fréquemment des sujets. Étant donné que le sujet peut être inversé, nous avons évalué également l'apport des mêmes entités-R positionnées à droite du verbe. Ceci permet d'améliorer le rappel de 0,05, avec de forts scores de précision sauf pour les lieux.
En utilisant une grammaire de segments, la précision chute quelque peu, mais on augmente significativement le rappel, à près de 0,1 de plus pour les catégories considérées (224 bons candidats). Les groupes nominaux les acronymes, les nom propres (*_np*) ainsi que les entités-R incertaines (*_pers_bof*) ont une précision moyenne à gauche. Les EN sont, au contraire, généralement précises sauf les lieux dont la précision est nulle. L'ajout de candidats à droite contribue également à ramener des sujets de longue portée (+0,02) avec un très faible nombre d'erreurs. 91 nouveaux candidats apportés par ce modèle sont bons.

Les causes d'erreurs de la grammaire syntaxique sont diverses mais on peut en retenir les principales :

- Le tiret n'est pas considéré comme une frontière : si un GN se trouve dans une insertion entre deux tirets, il est sélectionné.
- La coordination : les verbes peuvent être coordonnés, et si le premier possède un GN comme objet, c'est ce candidat qui est sélectionné. En revanche, s'il n'y a pas de GN, la grammaire peut détecter le sujet des deux verbes, dans le modèle M1 comme M2.
- EN non rattachées : le chunking n'associe généralement pas un groupe, dont il a déjà trouvé une entité-R tête, à une entité nommée. Cette entité nommée qui peut dépendre d'un GNP est donc libre et peut être sélectionnée par la grammaire syntaxique.

Les silences peuvent s'expliquer par :

- La non prise en compte de la totalité des catégories : il reste encore près de 35% de catégories à analyser, bien qu'elles puissent être occupées par des compléments d'objets. En particulier, les pronoms et les groupes nominaux inversés (à droite) peuvent améliorer les résultats.
- Les imbrications complexes de segments non gérées par la grammaire inter-segment
- Les erreurs d'étiquetage ou absence de listes

Les résultats en f-mesure ne sont pas comparables à l'état de l'art (entre 0,8 et 0,9 pour la relation sujet), mais ce n'était pas notre objectif de concevoir un analyseur syntaxique compétitif. Cette évaluation nous semble très instructive du point de vue de la caractérisation quantifiée des catégories pouvant apparaître en position sujet, que nous n'avons trouvée nulle part dans la littérature, notamment parce que nous employons des catégories sémantiques.

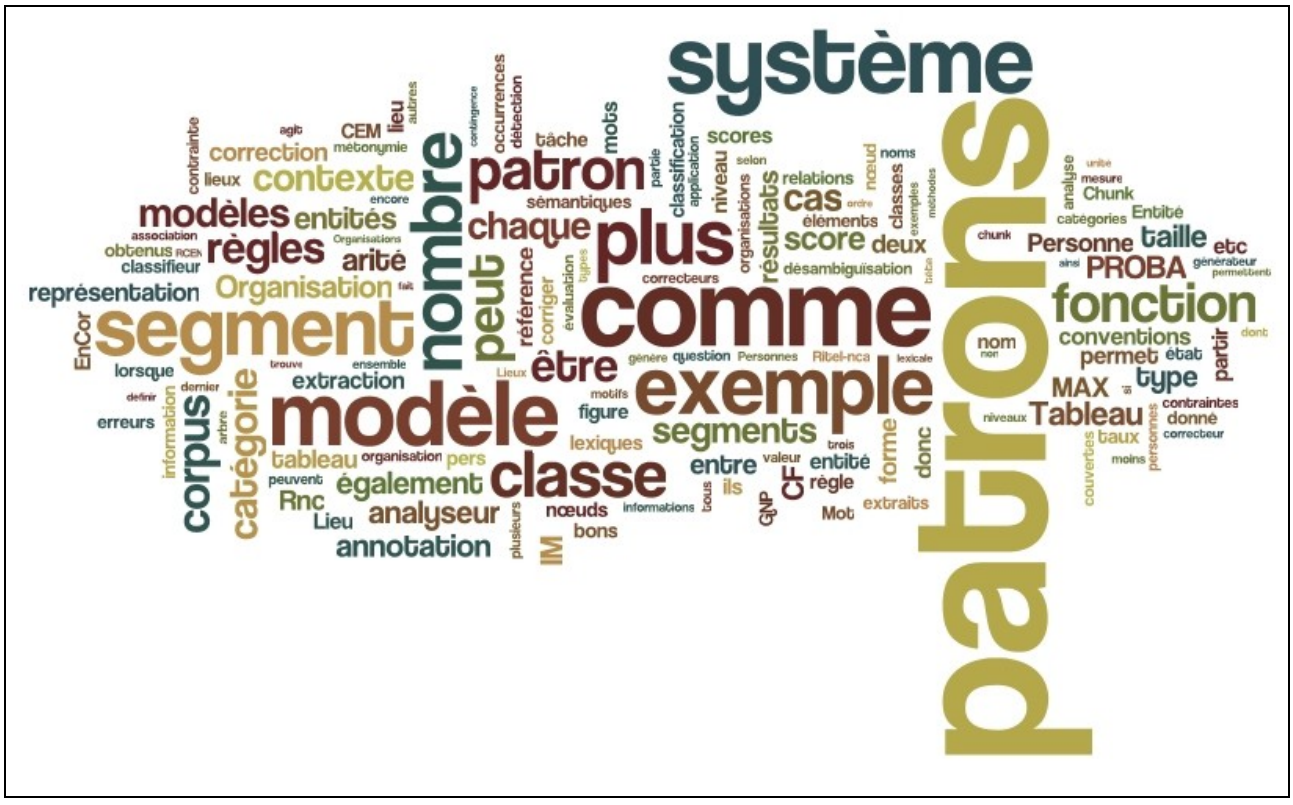
Si l'on s'en tient à cette représentation sémantique, cette caractérisation permet à quiconque intéressé par la détection de sujet de savoir que (en extrapolant à partir de ce verbe) les sujets sont réalisés à 10% par des groupes nominaux, à 10% par des pronoms et à 30% par des Npr ou expressions référentielles, à gauche du verbe. La gestion des sujets inversés peut permettre de détecter plus de 5% des candidats et la détection de sujets à longue distance (à gauche comme à droite) peut contribuer à hauteur de 10%. Ces résultats nécessitent évidemment d'être validés sur d'autres verbes ou d'autres formes verbales que celles que nous avons considérées. Il est fort probable que ces distributions changent selon les verbes, mais il est également possible que la distribution de verbes de même classe (citationnel) soit similaire.

Nous parvenons au terme des analyses que nous avons effectuées à partir des travaux effectués sur les relations inter-segment. Le prochain volet est consacré à l'analyse intra-segment, plus spécifiquement à l'extraction de patrons sémantiques au sein des segments et sera appliquée à la désambiguïsation des Entités Nommées.

CINQUIÈME VOLET

ANALYSE INTRA-SEGMENT APPLIQUÉE À LA DÉSAMBIGUISATION D'ENTITÉS NOMMÉES

Le modèle de représentation linguistique que fournit le segmenteur construit des espaces d'information de taille et de nature variable. Dans ce volet, nous chercherons à exploiter cet espace pour extraire des relations sémantiques pertinentes liées à une EN. Nous définirons plusieurs méthodes d'extraction de ces relations à partir de patrons. Ces patrons ne représentent pas directement des relations syntaxiques : l'ordre n'y est plus considéré comme un critère nécessaire, seule la co-occurrence de deux ou plusieurs éléments dans un segment prime. Nous chercherons plus spécifiquement à savoir dans quelle mesure ces patrons captent des régularités sémantiques, en évaluant leur potentiel de désambiguïsation d'EN. La désambiguïsation d'EN peut se concevoir comme un problème de classification dans lequel un système doit sélectionner la catégorie à laquelle est associée une EN dans un contexte donné, à partir d'un certain nombre d'informations. Nous utiliserons principalement le système de RCEN Rnc comme source d'information et la tâche consistera à le corriger. Le chapitre 10 est consacré à la description du système de correction et le chapitre 11 à l'évaluation.



10. Le système EnCor

Le système EnCor (Correction d'Entités nommées) se positionne en aval d'un système de RCEN dont il exploite les sorties pour le corriger dans un contexte donné. Il s'inscrit donc dans la problématique de l'adaptation et s'appuie sur les résultats des systèmes de RCEN Rnc et de chunking LoRit pour prendre des décisions. Nous avons vu (chapitre 5) que l'adaptation pouvait correspondre à de nombreuses tâches, comme de nouvelles définitions d'EN (métonymie), de nouveaux styles, domaines et genres de corpus. EnCor a été principalement conçu pour adapter le système Rnc à de nouvelles définitions (telles que fournies par les conventions d'annotation). Ce correcteur s'appuie sur le segmenteur que nous avons utilisé pour la détection de relations inter-segment : il extrait des patrons sémantiques à l'intérieur des segments qu'il exploite pour créer des règles de correction. Ces patrons sémantiques sont définis en fonction d'un niveau de représentation des éléments qui composent le segment, *i.e.* les chunks, les entités-R et les formes. Les règles sont représentées dans un formalisme transparent et lisible pour l'utilisateur. EnCor n'applique pas ces règles, il les génère. L'application des règles est effectuée par le même analyseur que nous avons utilisé pour définir la grammaire inter-segment et la grammaire syntaxique. Cette synergie des systèmes (segmenteur, analyseur et correcteur) permet par conséquent d'appliquer des règles manuelles et/ou des règles automatiquement induites de corpus. Pour comprendre le fonctionnement d'EnCor (9.2), nous présenterons donc d'abord l'analyseur (9.1).

10.1. L'analyseur

10.1.1. Principes

L'analyseur est hybride dans le sens où les règles qu'il applique peuvent être manuellement créées par un utilisateur ou induites automatiquement à partir de corpus. Les règles (et leurs lexiques associés) sont externalisées (dans des fichiers) et représentées dans un format spécifique. Elles sont exploitées comme paramètres par l'analyseur qui les traduit en opérations de parcours et de modification d'arbre.

Nous avons défini deux scénarios principaux dans lequel intervient cet analyseur. L'extraction de relation et la correction. Ces deux scénarios sont illustrés figure (10.1).

L'analyseur est utilisé plusieurs fois dans chaque scénario. Il est employé pour la pré-segmentation, la post-segmentation, l'analyse syntaxique et la correction. C'est pour l'appliquer à une telle diversité de tâches que nous l'avons conçu aussi générique que possible. Les scénarios partagent en commun un prétraitement en deux étapes : la pré-segmentation, qui a pour fonction de préparer le texte à l'application de la segmentation.

Le modèle que nous avons choisi pour représenter les règles sont des machines à état fini (MEF) que l'analyseur décompose état par état. Si la majorité des MEF suivent l'ordre de lecture, nous avons incorporé un mécanisme pour décider du bond (parcours dans l'arbre) à effectuer pour parvenir à l'état suivant : l'ordre de lecture est indépendant de l'ordre des états.

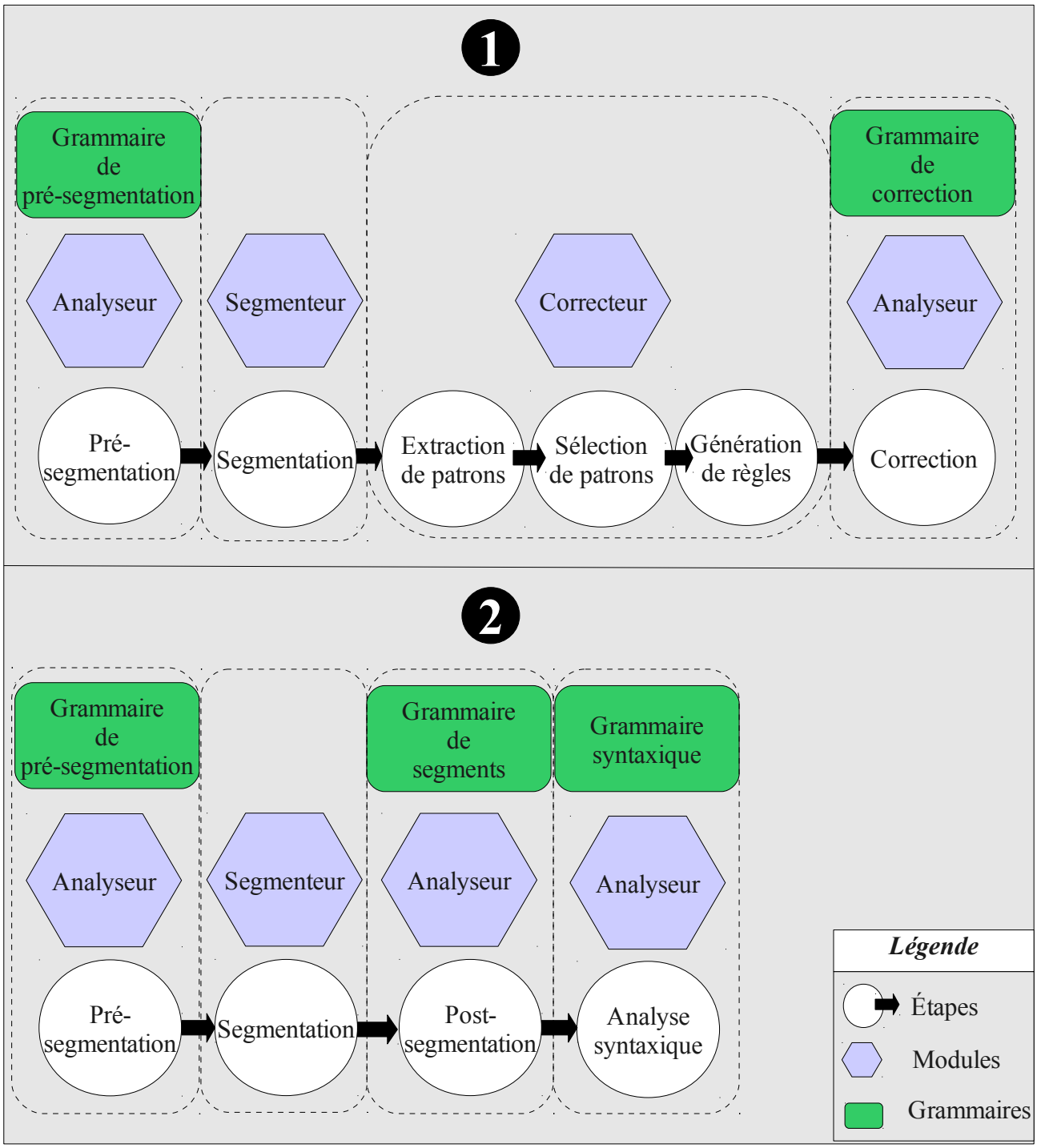


FIG 10.1 – Scénarios et modules de la chaîne de traitement

L'analyseur ne parcourt pas l'arbre en profondeur ni en largeur : comme l'arbre est fortement structuré, il se positionne directement sur les nœuds qui relèvent d'une classe donnée (Mot, Entité-R, Chunk, Segment), détectés à partir d'un index construit lors de l'extraction de l'arbre (pour les mots, entités-R, chunks) et de la segmentation (pour les segments). L'ordre par défaut d'application des règles est fonction de la complexité des nœuds : *Mot >Entité-R >Chunk >Segment*. L'analyseur teste chaque règle sur l'ensemble des nœuds appropriés. Il réalise trois fonctions primordiales :

(a) La validation d'un état consiste à traduire une expression (la règle) en un ensemble de contraintes minimales : la valeur d'un trait d'un nœud rencontré est comparée à la valeur attendue par une contrainte. La structure de ces contraintes est décrite en (10.1.3).

(b) La recherche du nœud suivant dépend de la classe de la contrainte du (ou des) prochain(s) état(s) de l'automate d'une règle ainsi que du bond à effectuer. L'algorithme de recherche du prochain nœud est décomposé en fonction des combinaisons possibles entre le nœud courant et le nœud du prochain état : soit ils sont identiques (Entité-R/Entité-R, etc.) soit ils sont différents (Entité-R/Mot, etc.). Le bond indique le nombre de fois que l'opération doit être répétée et le parcours s'arrête en fin de segment.

(c) L'application des actions d'une règle consiste essentiellement à appeler des méthodes sur des nœuds en fonction de paramètres. Certaines actions sont très simples, comme la modification d'un trait à un nœud donné ; d'autres, comme l'ajout de segment, supposent l'analyse du contexte du nœud pour décider du mode de rattachement à effectuer.

10.1.2. Structure des lexiques

Le modèle que nous avons choisi pour représenter les lexiques est une structure de traits. Les lexiques, comme les règles, sont structurés en quatre classes.

Une unité lexicale correspond à un noyau de traits exprimés sous forme de contraintes. Ce n'est pas nécessairement un mot ou une catégorie. Chaque contrainte est un triplet de la forme *<attribut, valeur, opérateur>*. Le type d'opérateur possible est P (positif, par défaut) ou N (négation de la contrainte). C'est une information que peut indiquer l'utilisateur pour définir des unités par exclusion (par exemple tout chunk qui ne soit pas de type GNP). Toutes les entrées possèdent un trait commun, le type, permettant de distinguer leurs classes Mot, Entité-R, Chunk et Segment. Le tableau (10.1) illustre la contrainte servant à déterminer la classe d'un chunk.

Attribut	Valeur	Opérateur
Type	*Chunk*	*P(ositif)*

Tableau 10.1 – Exemple de contrainte exprimée en 3-uple

Dans l'exemple, les attributs et les valeurs sont des constantes, mais elles peuvent être exprimées au moyen d'expressions régulières qui sont interprétées par le moteur lors de l'unification. Ceci permet de créer des ensembles lexicaux (listes de verbes, de signes de ponctuation, de conjonctions, etc.) pour les mots par exemple, ou des ensembles de catégories pour les entités-R, qui sont séparément définis comme variables pour être réutilisables dans plusieurs contraintes. Par exemple, la variable identifiant les catégories de personne peut être employée comme valeur d'une contrainte sur le trait HEAD (tête) d'un nœud de type Entité-R (tableau 10.2).

Attribut	Valeur	Opérateur
Type	*Entité*	*P(ositif)*
HEAD	*$_personne*	*P(ositif)*

Tableau 10.2 – Exemple de contrainte sur les entités-R Personne, où $_personne='(_pers|_pers_bof|_fp|_ap|_fap)'

Cette méthode permet d'ajouter autant de traits que nécessaire. Les traits que nous avons définis pour chaque classe sont alignés sur les informations fournies à l'analyseur.

Pour la classe Mot, il s'agit de la forme et des catégories morphosyntaxiques, qui sont récupérées par l'analyseur lors de l'analyse de l'arbre d'entrée. En corpus, les mots se présentent sous la forme d'une liste d'information dont la première est la forme, comme illustré en (237).

(237)Baudelaire|_~N|_~ms|_~Hum|~Npropr

Pour la classe Entité-R, les traits sont la tête et la forme de la tête, comme illustré en (238)

(238) <_pers> <_nom> Baudelaire|_~N|_~ms|_~Hum|~Npropr </nom></_pers>

La tête correspond à l'entité-R de plus haut niveau. Lorsque, comme dans cet exemple, il existe des filles uniques (l'entité-R *_nom*), elle est stockée à part comme alternative.

Pour la classe Chunk, il s'agit des mêmes informations, auxquelles sont ajoutées celles qui sont fournies par LoRit, le sous-type de chunk (GN, GNP, etc.), et d'autres propres à ce sous-type comme la préposition d'un GNP ou la polarité d'un verbe, comme on peut le voir en (239).

(239)<chunk_GNP_2_de><_prep> de </_prep><_pers> <_nom> Baudelaire|_~N|_~ms|_~Hum|~Npropr </nom></_pers></GNP>

Dans l'exemple, la tête est indiquée sous forme de numéro de fille, qui est identifiée lors de l'analyse du corpus.

Enfin, concernant la classe Segment, il s'agit des informations construites par le segmenteur : la taille, la nature des frontières, comme illustré en (240).

(240)<SEG_,_,_1><chunk_GNP_2_de><_prep> de </_prep><_pers> <_nom> Baudelaire|_~N|_~ms|_~Hum|~Npropr </nom></_pers></GNP></SEG>

Les séquences de formes, d'entités-R et de chunks qu'ils contiennent sont également stockées comme attributs pour permettre de définir des expressions régulières comme contrainte d'une unité lexicale (recherche de segment contenant un verbe).

Ces quatre exemples d'unités lexicale sont reproduits sous forme d'objets dans la figure (10.2).

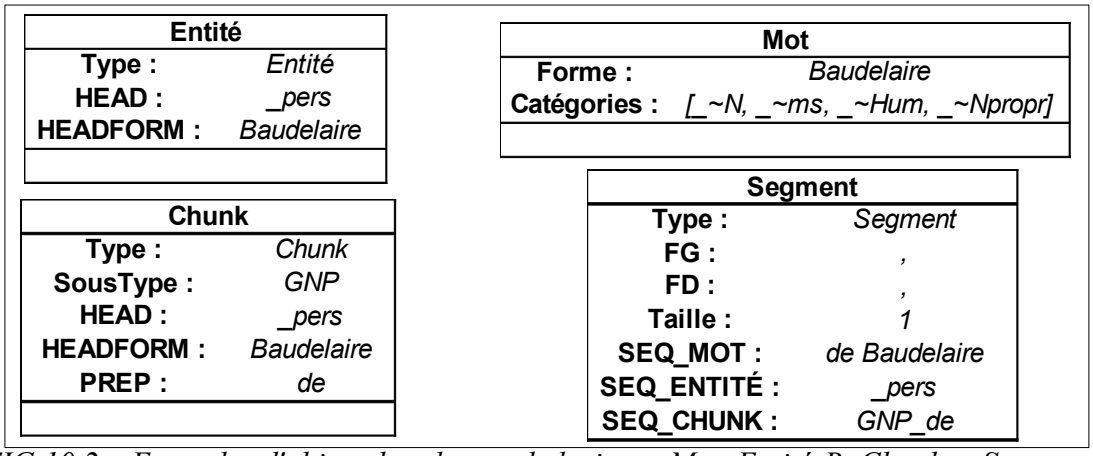


FIG 10.2 – Exemples d'objets des classes de lexiques Mot, Entité-R, Chunk et Segment

Chaque unité lexicale est un objet que l'utilisateur peut créer en manipulant des méthodes de classe Lexique. Les méthodes sont utilisées pour la création de l'unité et pour son utilisation et héritées par les classes Mot, Entité-R, Chunk et Segment.

Pour définir certaines listes de mots, nous avons utilisé les lexiques Morphalou et Delac : ceci nous a permis d'obtenir des listes de verbes, de participes ou encore de gérondif, extraites puis normalisées pour être employées dans les lexiques. C'est dans les lexiques que sont définies les frontières, ce qui signifie que l'on peut modifier leur nature sans que le fonctionnement du système soit affecté.

La conception d'unités lexicales suppose évidemment une certaine familiarité avec la forme que peuvent prendre les mots ou catégories que l'on cherche à définir.

10.1.3. Structure des grammaires

Le forte structuration des lexiques permet d'alléger la définition des règles des grammaires. Les grammaires sont externalisées et déclaratives, définissent à la fois les conditions et les actions. Les conditions correspondent à une liste ordonnée de contraintes invoquant des unités lexicales. Ces contraintes constituent les états d'un automate de règle évalué par l'analyseur et sont représentés sous la forme d'un quadruplé <attribut, valeur, sous-règle, position>. L'attribut définit la classe de lexique, la valeur définit l'unité lexicale et la position définit le bond à effectuer pour parvenir à l'état. Les sous-règles agissent comme des contraintes supplémentaires que ne permet pas d'exprimer une unité lexicale à elle-seule. Il s'agit principalement de contraindre l'environnement du nœud courant en évaluant les propriétés des nœuds adjacents. Elles permettent également de vérifier des contraintes sur des nœuds précédemment validés par la règle (comme pour une liste d'EN).

Les actions correspondent à des opérations génériques sur les nœuds. Elles sont de deux types : elles modifient la structure de l'arbre en déplaçant des nœuds ou modifient certaines valeurs de nœuds. Une règle formée de deux états, dont le premier est une initiale de la forme *M.,* et le second, un mot inconnu à capitale, pourra modifier l'étiquette en Personne (*_pers*) par exemple. Les actions de rattachement sont distinguées en fonction des classes de nœuds à déplacer. Certains, comme les segments, supposent une analyse fine du contexte, épargnée au concepteur de règle.

Comme les lexiques, chaque classe de règle fait partie d'une classe Règle, qui définit des méthodes permettant de créer ou d''utiliser les objets de classe règle.

Nous avons déjà décrit les applications de ces grammaires pour l'analyse des relations inter-segment. Une grammaire a été également créée en prétraitement à la pré-segmentation, pour limiter les erreurs du segmenteur. Par exemple, certains règles de classe Mot associent des nombres décimaux complexes, certaines abréviations spécifiques au corpus de presse trop fréquentes pour être ignorées (*div.* pour *divers gauche* par exemple dans le cas de partis politiques). Enfin, le segmenteur utilise certaines règles pour détecter les frontières.

10.2. Le correcteur EnCor

Le correcteur est utilisé dans le cas du scénario de correction. Il regroupe plusieurs modules que sont l'extracteur de patrons, le classifieur et le générateur de règles.

10.2.1. L'extracteur

L'extracteur permet de définir la forme des patrons, en sélectionnant le niveau de représentation approprié et son arité (le nombre d'éléments constituant un patron). Ce module s'appuie sur les segments comme fenêtre d'extraction.

Lorsque l'arité est fixée à 1, chaque nœud-fille du segment constitue un patron. Les patrons peuvent s'appuyer sur un niveau de représentation donné : Chunk, Entité-R ou Mot. À titre d'illustration, la figure (10.3) décline les caractéristiques internes de chaque segment de l'exemple (241) en fonction du niveau de représentation.

(241)il y a près de cinquante ans déjà , Jacques Monod rappelait au colloque de Caen que 50 pour cent du chiffre d' affaires de la société américaine Du Pont de Nemours provenait de la commercialisation de produits inconnus dix ans plus tôt .

Segment 1								
Chunk	-	-	-					
Entité	<_pres>	<_annee_dur>	<_adv>					
Forme	il y a	50 ans	déjà					
Segment 2								
Chunk	-	-	GNP_au	GNP_de				
Entité	<_pers>	<_action>	<_subs>	<_loc>				
Forme	Jacques Monod	rappelait	colloque	Caen				
Segment 3								
Chunk	GN	GNP_de	-	GNP_de	-	GNP_de	GNP_de	-
Entité	<_subsnn>	<_org>	<_pers>	<_loc>	<_action>	<_subs>	<_subs>	<_time>
Forme	chiffre d'affaires	société américaine	Du Pont	Nemours	provenait	commercialisation	produits	dix ans plus tôt

FIG 10.3 – Tableaux des niveaux de représentation des segments de l'exemple (241)

Comme on le constate tous les nœuds filles d'un segment n'ont pas été associés dans des chunks (présentatif *il y a*). Si le chunk n'existe pas, une catégorie par défaut est associée au nœud.

Cette représentation nivelée de la phrase permet de concevoir des modèles qui combinent les informations de plusieurs niveaux. L'extracteur crée trois modèles :

- *Le modèle CF* : il combine les niveaux Chunk et Mot.
- *Le modèle CE* : il combine les niveaux Chunk et Entité-R.
- *Le modèle CEM* : il combine les niveaux Chunk et Entité-R, en substituant les entités-R *substantif*, *action* et *adjectif* par les Mots correspondants, les verbes étant lemmatisés. L'existence de ce dernier modèle est motivée par l'hypothèse que ces classes contiennent régulièrement des informations sémantiques pertinentes qui seraient autrement masquées.

Le tableau (10.3) illustre les patrons extraits du segment 2 de l'exemple (241) pour chaque modèle.

	P1	*P2*	*P3*	*P4*
CF	**Jacques Monod**	**rappelait**	**GNP_au/colloque**	**GNP_de/Caen**
CE	**pers**	**action**	**GNP_au/subs**	**GNP_de/loc**
CEM	**pers**	**rappeler**	**GNP_au_colloque**	**GNP_de/loc**

Tableau 10.3 – Exemples de patrons extraits du segment 2 en fonction des modèles.

Alors que le modèle CEM cherche à optimiser les informations détenues par chaque élément, le modèle CE est le plus générique. Quant au modèle CF, il peut être plus précis lorsque les entités-R sont erronées (comme *Dupont Nemours* dans le segment 3 de l'exemple 241).

Pour construire des patrons d'arité supérieure à 1, l'algorithme calcule l'ensemble des combinaisons de patrons d'arité 1 au sein des segments. Il peut calculer

(a) les n-grammes (ou encore segments répétés ; [Lebart & Salem, 1994]) ou

(b) les motifs

(a) Les n-grammes correspondent aux séquences ordonnées possibles au sein d'un segment en fonction de l'arité, qui doit être inférieure ou égale à la taille du segment (dans le cas échéant, aucun patron n'est extrait). Par exemple, Les n-grammes d'arité 2 extraits du segment 2 de taille 4 de l'exemple (241) sont au nombre de trois (tableau 10.4) par modèle.

	P1	*P2*	*P3*
CF	**Jacques Monod|rappelait**	**rappelait|GNP_au/colloque**	**GNP_au/colloque|GNP_de/Caen**
CE	**pers|action**	**action|GNP_au/subs**	**GNP_au/subs|GNP_de/loc**
CEM	**pers|rappeler**	**rappeler|GNP_au_colloque**	**GNP_au_colloque|GNP_de/loc**

Tableau 10.4 – N-gramme d'arité 2 du segment 2 de l'exemple 241

Le nombre de n-grammes croît linéairement selon l'arité (A) et de la taille (T) du segment.

$$Ngram = T - A + 1$$

(b) Les n-grammes sont des cas particuliers des motifs dont l'ordre séquentiel n'est pas contraint. Les motifs correspondent aux combinaisons uniques non-ordonnées au sein d'un segment. Par exemple, pour le segment 2 de l'exemple (241), nous souhaitons, pour chaque modèle, générer les patrons décrits dans les tableaux (10.5) à (10.7) : il s'agit de toutes les combinaisons possibles classées par arité.

Arité	#	CF
2	P1	Jacques Monod|rappelait
	P2	Jacques Monod|GNP_au/colloque
	P3	Jacques Monod|GNP_de/Caen
	P4	rappelait|GNP_au/colloque
	P5	rappelait|GNP_de/Caen
	P6	GNP_au/colloque|GNP_de/Caen
3	P7	Jacques Monod|rappelait|GNP_au/colloque
	P8	Jacques Monod|rappelait|GNP_de/Caen
	P9	rappelait|GNP_au/colloque|GNP_de/Caen
4	P10	Jacques Monod|rappelait|GNP_au/colloque|GNP_de/Caen

Tableau 10.5 – Motifs générées pour le modèle CF (segment 2 de l'exemple 241)

Arité	#	CE
2	P1	pers|action
	P2	pers|GNP_au/subs
	P3	pers|GNP_de/loc
	P4	action|GNP_au/subs
	P5	action|GNP_de/loc
	P6	GNP_au/subs|GNP_de/loc
3	P7	pers|action|GNP_au/subs
	P8	pers|action|GNP_de/loc
	P9	action|GNP_au/subs|GNP_de/loc
4	P10	pers|action|GNP_au/subs|GNP_de/loc

Tableau 10.6 – Motifs générées pour le modèle CE (segment 2 de l'exemple 241)

Arité	#	CEM
2	P1	pers|rappeler
	P2	pers|GNP_au_colloque
	P3	pers|GNP_de/loc
	P4	rappeler|GNP_au_colloque
	P5	rappeler|GNP_de/loc
	P6	GNP_au_colloque|GNP_de/loc
3	P7	pers|rappeler|GNP_au_colloque
	P8	pers|rappeler|GNP_de/loc
	P9	rappeler|GNP_au_colloque|GNP_de/loc
4	P10	pers|rappeler|GNP_au_colloque|GNP_de/loc

Tableau 10.7 – Permutations générées pour le modèle CEM (segment 2 de l'exemple 241)

Les tableaux ne font pas figurer les patrons d'arité 1 (voir tableau 10.2). Bien évidemment, l'arité est égale à la taille du segment, il ne peut y avoir qu'un seul patron (le segment selon le modèle de représentation). Le nombre de motifs d'arité A correspond au nombre de combinaisons de A éléments dans un ensemble à n éléments (C_n^A). Le nombre total de motifs est donc $C_n^1+C_n^2+...+C_n^n=2^n-1$, le nombre de parties d'un ensemble à n éléments lorsqu'on exclut la partie vide. Le nombre de permutations à calculer dépend de la taille du segment : il devient rapidement prohibitif lorsqu'il est calculé dans des segments longs (figure 10.4). Le tableau (10.8) illustre le nombre de patrons obtenus en fonction de l'arité A et de la taille du segment T, c'est-à-dire en fonction des cas de figure que l'on peut rencontrer. Le total pour chaque taille correspond à la suite obtenue par 2^n-1.

	T=1	T=2	T=3	T=4	T=5	T=6
A=1	1	2	3	4	5	6
A=2	0	1	3	6	10	15
A=3	0	0	1	4	10	20
A=4	0	0	0	1	5	15
A=5	0	0	0	0	1	6
A=6	0	0	0	0	0	1
Total	1	3	7	15	31	63

Tableau 10.8 – Nombre de patrons obtenus par permutation selon l'arité et la taille des segments

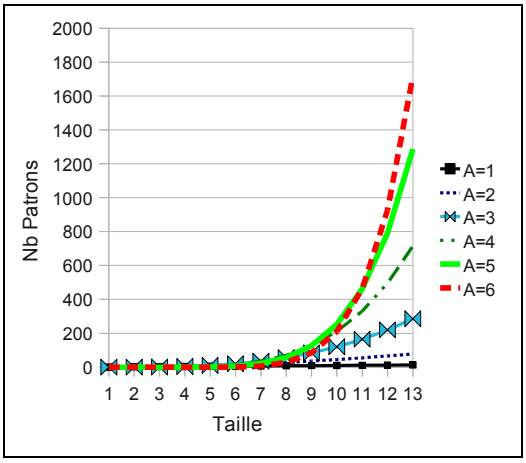


FIG 10.4 – Courbe d'évolution du nombre de patrons obtenus par les motifs

Pour les motifs, nous avons limité l'arité à 3 et la taille de segment à 7, afin de ne pas dépasser l'extraction de 50 patrons par segment.

Ces méthodes d'extraction ont été proposées à titre exploratoire et nous n'avons pas pu toutes les étudier en détail. Il faut noter que si le nombre d'extractions peut être grandement prohibitif, le nombre de patrons-types (et non leurs occurrences) est largement moindre, bien que difficilement exploitable par un être humain. Il convient donc de filtrer les patrons types extraits en tenant compte de leur fréquence en corpus comme mesure de base. La méthode de score qui sera employée sera fortement dépendante du type d'application envisagé. L'application à laquelle nous avons destiné le système EnCor est la désambiguïsation des EN. Les patrons (peu importe la méthode d'extraction), seront évalués en fonction de leur capacité discriminante vis-à-vis d'une classe d'EN donnée. Les n patrons obtenant les meilleurs scores vis-à-vis d'une classe d'EN, pourront être évalués manuellement. L'espoir que nous entretenons est que ces patrons capteront des propriétés du corpus, plus spécifiquement, des propriétés sémantiques de l'usage de ces EN en corpus.

10.2.2. Le classifieur

Le classifieur a été destiné à la tâche de désambiguïsation : il est chargé de sélectionner les patrons les plus significativement associés à une classe d'EN. Il est paramétré pour une méthode d'extraction d'arité 1 et les classes retenues sont *Personne*, *Organisation* et *Lieu*. Les classes peuvent correspondre à plusieurs entités-R ; par exemple, la classe Personne peut être réalisée par une entité-R *_pers* ou une entité-R *_fp*, qui contient une personne. Nous avons choisi de regrouper

plusieurs entités-R dans un lexique représentatif de la classe[48]. Elle a été établie d'après notre expérience de la classification opérée par le système Rnc. Une autre alternative consisterait à classer les patrons par rapport à chaque entité-R indépendamment et à n'effectuer le regroupement qu'après la classification, comme nous l'avions fait pour l'évaluation de la détection de relation sujet. On pourrait également envisager de caractériser plus finement les entités-R en fonction de la structure de l'arbre auquel elles appartiennent. Par exemple, on pourrait distinguer les entités-R *_pers* qui comportent un nom et un prénom de celles qui ne contiennent qu'un nom, en faisant l'hypothèse que cette dernière configuration a de plus fortes chances d'être erronée. Ce sont autant de manières de superviser la classification. En l'état actuel, le classifieur n'est véritablement supervisé que par le regroupement d'entités-R en classe, ses entrées comportent donc des erreurs.

Le classifieur sélectionne tous les segments du corpus comportant au moins l'une de ces EN et comptabilise les patrons extraits en fonction du niveau de représentation. Dans ce cas (arité=1), un patron correspond à un nœud-fille apparaissant dans un segment (sauf l'EN recherchée). Par exemple, le patron *GNP_de/_loc* du modèle CEM apparaît 6411 fois en co-ocurrence avec une entité-R de type *Lieu*, 2604 fois avec une *Organisation* et 3814 avec une *Personne*.

À partir de ces données, deux scores d'association d'un patron sont calculés pour chaque classe : la probabilité de cooccurrence entre un patron et une classe d'EN donnée (PROBA) et l'information mutuelle (IM) :

$$PROBA(EN \mid Patron) = \frac{P(EN, Patron)}{P(Patron)}$$

$$IM(EN,\ Patron) = P(X=EN, Y=Patron) \times \log \frac{P(X=EN, Y=Patron)}{P(X=EN) \times P(Y=Patron)}$$

La probabilité doit permettre de détecter des associations <*Patron, EN*> fréquentes et l'IM, qui caractérise la dépendance entre ces deux variables, des associations significatives pas nécessairement fréquentes.

Ces scores permettent de prédire la classe d'EN la plus probable vis-à-vis d'un patron donné. Trois scores globaux sont ensuite calculés afin de prendre en compte l'ensemble des patrons contenus dans le segment :

- la moyenne des scores des patrons (Mean) pour tenir compte du nombre de patrons dans le segment,
- le produit des scores (Prod), pour atténuer l'importance de patrons fréquents et communs aux différentes classes et
- le score du meilleur patron (Max)

Pour exemple, le score PP (Produit de Probabilités) d'un segment pour la classe Personne, se calcule à partir du produit des probabilités de ses patrons :

48 Les entités-R utilisées sont les suivantes :
Personne={_pers, _pers_comp, _pers_bof, _fap, _fp, _ap}
Organisation={_org, _org_gvt, _org_peup, _org_prob, _org_sport, _org_div, _org_medias, _org_pol, _org_relig, _org_univ, _prod}
Lieu={_loc, _ville, _pays, _montagne, _fleuve, _province, _lieu_remarquable, _departement, _quartier, _etats, _royaume, _territoire, _comte, _emirat, _republique, _ile, _atoll, _lagune, _littoral, _peninsule, _prequile, _archipel, _cap, _temple, _culte, _chateau, _palais, _desert, _foret, _grotte, _musee, _oasis, _plage, _plaine, _plateau, _principaute, _seuil, _site, _theatre, _vallee, _voie, _adresse, _coordonnee}

$$PP(Personne) = \prod_{1}^{n} PROBA(Personne|Patron_i)$$

Six scores différents ont ainsi été expérimentés pour chaque type de modèle de représentation, vis-à-vis de chacune des trois classes, Personne, Lieu et Organisation. À partir de ces différents modèles de mesure, le classifieur sélectionne la classe dont le patron, la somme de patrons ou la moyenne de patrons, est la plus forte.

10.2.3. Le générateur de règles

Le générateur de règles a été intégré pour traduire les patrons automatiquement extraits dans le format de règles interprété par le moteur. Il a été paramétré pour la désambiguïsation, mais on peut envisager, comme pour le classifieur, de l'utiliser dans d'autres tâches, comme le rattachement syntaxique. Dans le cadre de la tâche de désambiguïsation, le type d'action ciblé est systématiquement le même : la modification de l'étiquette du nœud-cible. Le générateur récupère en entrée une des matrices générées par le classifieur (selon le modèle de représentation et de score), qui indique en colonne, les scores obtenus pour chaque type d'entité nommée et en ligne, les patrons sélectionnés. Le tableau (10.9) indique les scores obtenus pour les 10 meilleurs patrons obtenus pour la classe Personne avec le modèle CF, en fonction du score Proba_Max, sur le corpus de développement.

Patron	*Personne*	*Organisation*	*Lieu*
jean	0,536	0,000	0,000
olivier	0,532	0,028	0,032
noter	0,521	0,000	0,000
jacques	0,463	0,000	0,001
expliquer	0,460	0,001	0,001
entourer	0,447	0,028	0,032
dit	0,429	0,002	0,001
souligner	0,397	0,002	0,002
robert	0,381	0,028	0,032
confier	0,377	0,003	0,002

Tableau 10.9 – Scores obtenus par les patrons CF en fonction de la classe d'EN

Deux types de patrons peuvent être identifiés : des prénoms mal normalisés (*jean*, *olivier*, etc.) et des verbes, dont une partie de verbes de citation (*expliquer*, *dit*, *confier*, etc.). Le générateur traduit ces patrons en règles en créant les entrées des grammaires et des lexiques selon le modèle. Dans ce cas (modèle CF), ce sont des unités lexicales et des règles de classe Mot. Les règles sont composées de deux états, le premier définissant les alternatives pour la classe d'EN, le second, le patron. L'action de la règle consiste à modifier l'attribut tête de l'EN identifiée dans le premier état.

Le générateur est le dernier composant du correcteur, ses règles sont ensuite appliquées par l'analyseur. Nous évaluons les performances de ce correcteur dans le chapitre suivant.

11. Adaptation d'un système de RCEN

Pour première expérience d'acquisition automatique de relations sémantiques, nous avons cherché à savoir si le système EnCor permettait d'extraire des patrons sémantiquement discriminants. À travers ce système d'extraction, nous cherchons à évaluer la dualité entre la catégorie sémantique d'une expression linguistique et les relations auxquelles elle est associée. Par exemple, un patron tel que *naît en 1982* aura fort probablement un sujet de type *Personne*. Nous ne disposons pas d'analyseur en dépendance mais d'un système d'extraction de patrons qui caractérise les EN en fonction des patrons sémantiques qui leur sont associés au sein d'un segment. Les segments retenus seront donc des segments simples (de taille 1) et seront exclus de l'évaluation. L'évaluation du système porte sur la problématique de l'adaptation : le classifieur doit corriger les erreurs du système Rnc. Quatre paramètres peuvent justifier la nécessité d'une adaptation :

- le système Rnc est développé pour la détection d'EN sur des corpus oraux ; nous l'évaluerons sur le corpus de presse écrite.
- Le système Rnc n'est pas développé pour traiter spécifiquement des conventions d'écriture (le style) du corpus de presse.
- Le corpus de presse contient des EN inconnues au système Rnc, qui peuvent influencer ses choix de catégorisation.
- Le système Rnc ne gère que certains phénomènes de métonymie.

Il serait raisonnable de considérer que le système Rnc est suffisamment robuste pour gérer ces paramètres de variation. Nous chercherons donc plus spécifiquement à évaluer la capacité du système Encor à l'adapter à de nouvelles conventions de définition d'EN. Comme décrit précédemment (chapitre 5), ces conventions ont évolué et sont encore discutées. Conformément à ce que nous avions proposé, nous établirons des conventions en fonction du contexte d'application, qui concerne l'extraction d'information pour un système de Question-Réponse, en l'occurence Ritel (11.1). Nous décrivons les résultats de l'évaluation en (11.2) et les patrons extraits par le système EnCor en (11.3).

11.1. Évaluation

11.1.1. Conventions d'annotation

L'annotation a été réalisée par une interface d'annotation web : en sélectionnant une EN dans le texte (figure 11.1), l'annotateur pouvait choisir une classe, grâce à un bouton (en vert), ou corriger une annotation précédente (en rouge) ; chaque EN était alors surlignée par une couleur distinctive.

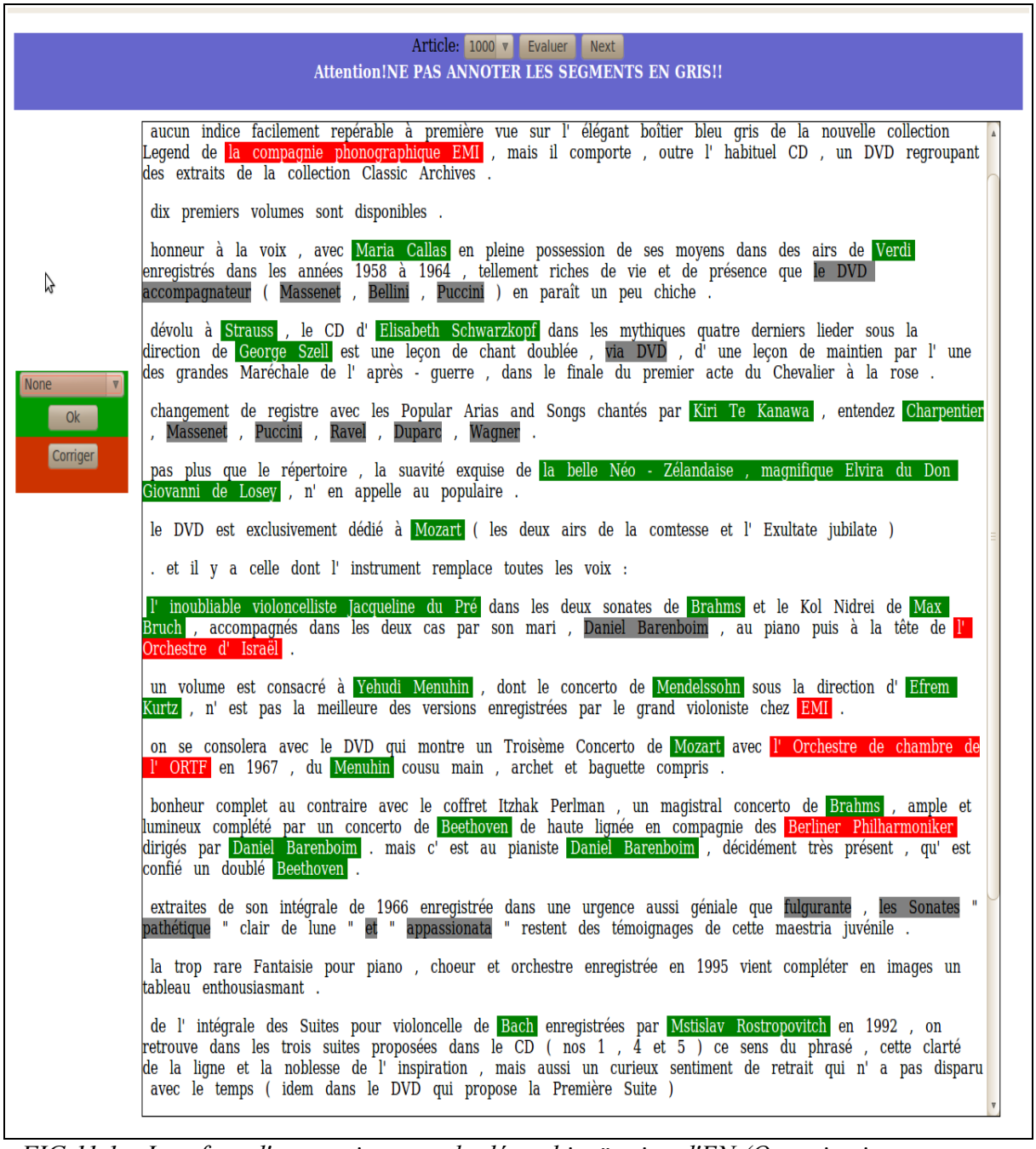


FIG 11.1 – Interface d'annotation pour la désambiguïsation d'EN (Organisations en rouge, Personnes en vert, Lieux en bleu, Segments simples en gris)

Le corpus de presse est divisé en corpus de développement, à partir duquel nous avons extrait les patrons (17 millions de mots) et en corpus de test (5,5 millions de mots, 10 000 articles). 200 articles de presse ont été annotés pour obtenir plus de mille instances d'entités de chaque classe (plus exactement 1426 organisations, 1004 lieux et 1377 personnes).

Les conventions d'annotation sont établies sur des critères linguistiques, contextuels et guidés par le contexte d'application.

Du point de vue de la forme, les conventions ont réduit la tâche de détection des EN au nom propre, avec ou sans majuscule, en utilisant notamment les titres, fonctions et déterminants pour délimiter les bornes de l'EN : ces bornes sont systématiquement exclues. Les expressions référentielles (comme *le président*) ont donc également été exclues. Néanmoins, lorsque certains éléments pouvaient être considérés comme constitutifs de la dénomination d'une EN, ils ont été inclus, ce qui distingue l'exemple (242) de l'exemple (243).

(242) l'<org> université de Poitiers </org>
(243) le maire de <lieu> Poitiers </lieu>

Dans ces exemples, c'est la dénomination globale qui l'emporte : *l'université de Poitiers* est une dénomination, *le maire de Poitiers* est la composition d'une fonction et d'un lieu. *Poitiers* en (243) pourrait être considéré comme une organisation, ou une entité geo-politique. Il est considéré comme un lieu, parce qu'il est la réponse à la question exprimée en (244).

(244) De quelle ville Alain Claeys est-il maire ?

Cette convention traduit directement notre préoccupation à inscrire le système de RCEN dans le cadre d'un SQR. Plus particulièrement, les EN sont considérées selon le rôle qu'elles jouent en contexte dans le cadre d'un scénario d'extraction d'information exprimé par une question. En (242), le contexte n'est pas suffisant pour déterminer à quelle question l'occurrence de l'EN est susceptible de répondre. La difficulté majeure consiste donc à identifier des critères fiables pour résoudre les cas où le type d'une EN diffère de son rôle en contexte. Les conventions d'annotation privilégient dans ces cas l'interprétation contextuelle. Deux types de divergences ont été rencontrées :

- lorsque cette divergence était explicitée par un déclencheur immédiatement apposé (245)
- lorsqu'elle était due à une interprétation globale de la phrase, ou du contexte de l'article (246)

(245) c' est la mesure phare de la loi Perben du 9 mars sur la criminalité
(246) l' Italie s' oppose à une réforme du Conseil de sécurité de l' ONU

En (245), l'EN *Perben* a été exclue de l'annotation car elle relève du type *Loi*, bien qu'elle soit nommée d'après son fondateur. Cela signifie que ce contexte n'est pas jugé pertinent pour répondre à une question sur l'origine du nom de cette loi, mais plutôt sur des relations telles que la date (247) ou le sujet (248).

(247)De quand date la loi Perben ?
(248)Quel sujet la loi Perben traite-t-elle ?

L'exemple (246) est généralement décrit comme un cas de métonymie (Markert et al., 2007), pour rendre compte de la relation existant entre un lieu (*Italie*) et des individus, l'interprétation étant due

au verbe avec lequel l'EN est employée. Cet exemple ne désigne pas une personne comme les conventions d'annotation de métonymie de la campagne Semeval7 [Markert & Nissim, 2007] semblent l'indiquer à travers la catégorie « Loc-for-People » (il n'existe pas de catégorie « *Loc-for-Org* ») : il s'agit d'une organisation politique, dans ce cas très probablement le gouvernement. La question a laquelle peut répondre cette occurrence est illustrée en (249).

(249)*Quel pays/Qui* s'oppose à une réforme du conseil de sécurité de l'ONU ?

D'autres types d'organisations répondent à ce phénomène de métonymie, comme les équipes de sport (250).

(250)dans les autres rencontres disputées mercredi soir, <org>Auxerre</org> s' est imposé à <loc>Rennes</loc>

Dans cet exemple, *Auxerre* est une organisation qui répond à la question (251).

(251)*Qui/quel club* s'est imposé à Rennes mercredi ?

En revanche, *Rennes* est employé comme localisation (à opposer à *s'imposer à domicile*). Pourtant, cet exemple pourrait servir de réponse à la question illustrée en (252)

(252)Quelle équipe Auxerre a-t-elle vaincue mercredi ?

Remarquons qu'on trouvera difficilement **à qui s'est imposé Auxerre ?*, et que dans l'exemple (252), nous avons substitué le verbe par *vaincre*. Par conséquent, c'est parce que nous raisonnons sur le plan sémantique que nous nous permettons cette substitution. D'un point de vue linguistique, la question « correcte » contenant le verbe *s'imposer* sera une variante de *Où Auxerre s'est imposée mercredi ?* ; dans ce cas la catégorie Lieu convient. Évidemment, il serait plus intéressant qu'un système puisse inférer l'équivalence sémantique entre *s'imposer* et des verbes comme *vaincre*, *battre*, *gagner*, etc. Les SQR comme RITEL proposent dans ce genre de cas d'utiliser des dictionnaires de synonymes pour l'expansion de requêtes, mais ces expansions sont effectuées à partir de dictionnaires de synonymes et sur la base d'une équivalence mot-mot. Or, nous avons montré dans les chapitres précédents que le sens était mieux appréhendé en contexte et en prenant en compte l'environnement linguistique du mot-cible. L'équivalence mot-mot est un pis-aller d'autant qu'il n'existe pas de ressource sémantique comme FrameNet pour le français, qui nous permette d'associer deux structures prédicatives équivalentes. Nous considérons donc que nous atteignons ici la limite de ce qu'une extraction sémantique du contexte peut apporter : la réponse à une telle question, étant donné ce contexte, devra être résolue par d'autres moyens. Du point de vue du système RITEL, cela signifie que le DDR (descripteur de recherche ; cf. *infra* 6.1.1) doit disposer, d'une manière ou d'une autre, de ces équivalences. Mais comme ce DDR intervient dans toutes les phases majeures qui suivent l'analyse de la question, l'intégration optimale de telles connaissances nécessite d'être étudiée en détail.
Pour ce qui nous concerne ici, les lieux ont donc été annotés comme tels lorsque l'interprétation en contexte le justifiait (notions de localisation, destination, origine, etc.) L'EN *Florence*, pourra désigner une personne (253) ou un lieu (254) selon le contexte.

(253)ce n' est pas le moindre des mérites de l' essai d'Anton Brender et Florence Pisani
(254)une forte pluie commença à tomber sur la Toscane et Florence

11.1.2. Accord inter-annotateur

Nous avons effectué une première évaluation avec deux annotateurs pour déterminer l'ensemble de ces conventions et dégager les types de désaccords. Nous avons retenu 6 articles pour un total de 100 EN (moins de 3% du corpus de test). À partir des résultats, nous avons calculé le coefficient Kappa sur l'accord de détection (entre catégories *EN* et *NON_EN*). Pour cela nous avons compté le nombre de mots considérés par tel annotateur comme (faisant partie d'une) EN et le nombre de mots non considérés comme des EN. Le coefficient Kappa de Cohen (1960) utilisé (pour deux annotateurs) se calcule selon la concordance observée (Po) et la concordance aléatoire (Pe) :

$$K = \frac{Po - Pe}{1 - Pe}$$

La table de contingence obtenue pour la tâche de détection est indiquée dans le tableau (11.1) :

Annot1\Annot2	EN	NON_EN	Total
EN	135	6	141
NON_EN	1	2477	2478
Total	136	2483	2619

Tableau 11.1 – Table de contingence sur la tâche de détection pour l'annotation

Sans surprise, le Kappa est excellent (K=0,97), puisque le nombre de mots total biaise le calcul, mais c'est une première information sur l'accord inter-annotateur au niveau de la détection des EN.

Pour étudier l'accord inter-annotateur sur la classification, nous avons retenu 101 EN détectées par les annotateurs. Nous avons décomposé l'analyse en fonction des catégories. Les tables de contingence pour chaque catégorie sont illustrées dans les tableaux (11.2) à (11.4).

Annot1\Annot2	PERS	NON_PERS	Total
PERS	46	0	46
NON_PERS	2	53	55
Total	48	53	101

Tableau 11.2 – Table de contingence sur la tâche de classification de personnes pour l'annotation

Annot1\Annot2	LOC	NON_LOC	Total
LOC	19	1	20
NON_LOC	1	80	81
Total	20	81	101

Tableau 11.3 – Table de contingence sur la tâche de classification de lieux pour l'annotation

Annot1\Annot2	ORG	NON_ORG	Total
ORG	30	4	34
NON_ORG	1	66	67
Total	31	70	101

Tableau 11.4 – Table de contingence sur la tâche de classification d'organisations pour l'annotation

Les coefficients Kappa obtenus sont de 0,96 pour les Personnes, 0,93 pour les lieux et 0,88 pour les organisations, ce qui nous a paru satisfaisant pour annoter la totalité du corpus de test.

11.2. Résultats

11.2.1. Performances des systèmes

Pour guider nos analyses, nous présentons tout d'abord les résultats obtenus par notre système de référence, Ritel-nca. Le tableau 11.5 indique le nombre d'EN figurant dans le corpus de test (N), le nombre d'EN détectées par le système (ramenés) et ses performances pour chaque classe.

CLASSE	N	RAMENÉS	CORRECT	PRÉCISION	RAPPEL	FMESURE
LIEU	1004	1197	689	0,58	0,69	0,63
ORGANISATION	1426	892	532	0,60	0,37	0,46
PERSONNE	1377	1425	1092	0,77	0,79	0,78
TOUS	3807	3514	2313	0,66	0,61	0,63

Tableau 11.5 – Rappel, Précision et F-mesure du système Ritel-nca

886 des 3807 EN annotées n'ont pas été classées par ce système, soit près de 25%, parmi lesquelles 515 sont détectées comme noms propres. Mis à part ce dernier cas, les « erreurs » de détection affectent principalement la catégorie Organisation et s'expliquent pour les raisons suivantes : EN non retenues, mots inconnus, problèmes de normalisation du texte (suppression de majuscules, encodage), etc. La cause d'erreur majeure de classification est due aux divergences entre les conventions d'annotation et celles qu'est censé appliquer le système R : la gestion de la métonymie. Le cas de la métonymie Lieu → Organisation entraîne un faible rappel des Organisations (0,37) et une mauvaise précision pour les lieux (0,58), par rapport aux évaluations précédentes de ce système [Rosset et al., 2005].

Étant donné que les patrons que génère le système EnCor classent les EN à partir des entités-R fournies par le système Ritel-nca, l'évaluation a uniquement porté sur les EN détectées (en excluant également les noms propres non classés) : une tâche de désambiguïsation. Les résultats de ce dernier ont été recalculés et figurent dans le tableau (11.6).

CLASSE	N	RAMENÉS	CORRECT	PRÉCISION	RAPPEL	FMESURE
LIEU	824	1029	689	0,67	0,84	0,74
ORGANISATION	949	614	532	0,87	0,56	0,68
PERSONNE	1148	1269	1092	0,86	0,95	0,90
TOUS	2921	2912	2313	0,79	0,79	0,79

Tableau 11.6 – Rappel, Précision et F-mesure du système Ritel-nca sur les EN détectées

Le nombre de segments total de ce sous-corpus s'élève à 1712, réduisant le nombre de segments contenant au moins une personne détectée, à 943 (les lieux à 818 et les organisations à 659). Il faut également noter que 28% des segments contiennent plus d'une entité, ce qui peut poser des difficultés : si deux (ou plus) entités nommées ne sont pas de même type, le meilleur patron ne pourra en discriminer au plus qu'une seule.

Par degré d'importance, le score d'association (IM, PROBA) est la variable qui influence le plus les résultats, suivi par le niveau de représentation (CE, CF, CEM). Quant au calcul global du score (MAX, PROD, MEAN), il n'a qu'une faible influence : le choix du meilleur score d'association (MAX) équivaut globalement à calculer la moyenne ou le produit des scores de tous les patrons. Par conséquent, les diagrammes de la figure (11.2) font uniquement figurer les moyennes des scores en fonction de la mesure d'association (PROBA, IM). Les résultats sont présentés en fonction de la taille des segments pour chaque classe d'EN ; le nombre de segments par

taille figure sur les diagrammes (NS), ainsi que les résultats de Ritel-nca (R), à titre comparatif.

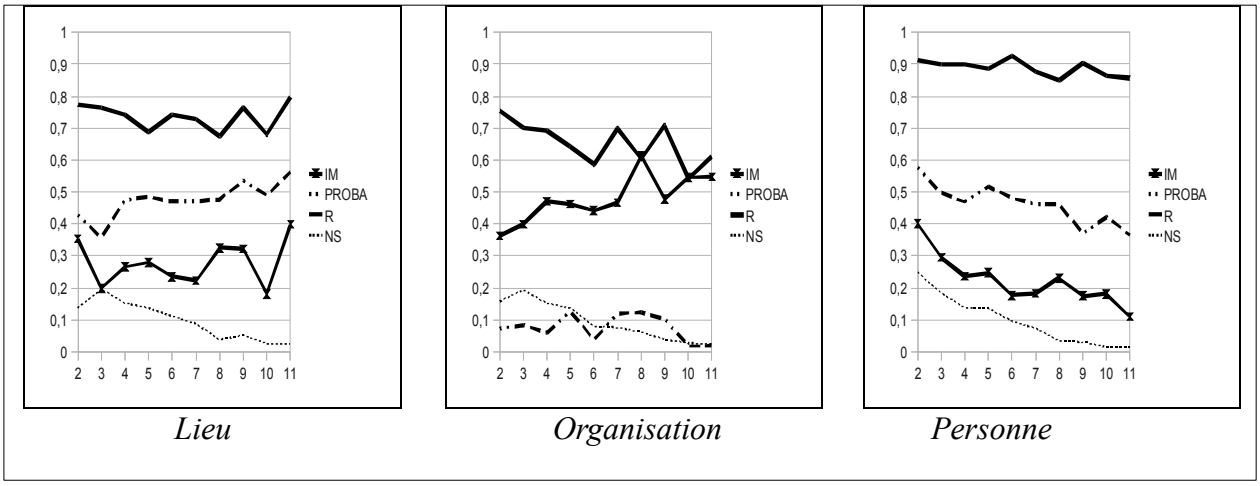


FIG 11.2 – F-mesure du système EnCor pour les Lieux, les Organisation et les Personne

Comme on peut l'observer, ces modèles ne rivalisent pas avec le modèle de référence (R). Individuellement, les meilleurs modèles atteignent 0,62 de F-mesure sur les Personnes, 0,55 pour les Lieux et 0,45 sur les Organisations. On peut retenir globalement que le score PROBA est plus approprié pour la classification des Lieux et des Personnes, alors que l'IM semble plus performante sur les Organisations. L'augmentation de la taille du segment semble avoir un impact négatif sur la classification des Personnes (plus un segment est long, moins les indices sont discriminants), mais elle est liée à une amélioration des modèles PROBA pour les Lieux et des modèles IM pour les Organisations.

Comme les modèles s'appuient sur l'annotation du système de référence, on peut supposer que les erreurs effectuées par ce dernier seront reproduites par les modèles et qu'une partie des erreurs des modèles sont dues au fait que l'acquisition des patrons s'est effectuée sur un corpus comportant des erreurs. Notons également que les modèles ne s'appuient sur aucune autre connaissance que le corpus de développement, ce qui n'est pas le cas du système Ritel-nca qui manipule des listes d'EN de personnes, de lieux, et d'organisation. La segmentation effectuée, qui limite l'espace de sélection de patrons et la nature des patrons-candidats dans chaque segment, joue également un rôle dans l'identification des patrons. Il nous reste par conséquent à analyser les patrons sélectionnés par chaque modèle pour savoir s'ils reflètent des usages typiques d'EN ou des usages spécifiques au corpus, soit à évaluer leur utilité dans une tâche de correction.

11.2.2. Correction du système de référence

Sachant que la mesure de score global a une influence minime sur les résultats, nous avons sélectionné tous les modèles MAX : ils extraient un patron (un chunk du segment) pour chaque EN. Par exemple, le modèle CEM_IM_MAX a classé correctement 17 instances de personnes grâce au patron *expliquer*, comme dans l'exemple (255) :

(255) " mon rôle est de bousculer la perception que les gens ont de Burberry ", *explique* Christopher Bailey

L'exemple (256) est une erreur que ce patron permet de corriger : R a classé *Maud* en Lieu.

(256)une sorte de tri sélectif qui " élimine les cellules mortes et rend la peau douce et satinée ", explique *Maud*

Les résultats de l'évaluation nous permettent d'assigner un taux de réussite à chaque patron. Chaque patron peut être associé à un taux de couverture : le nombre de fois qu'il a été sélectionné par le modèle comme discriminant, c'est-à-dire le nombre d'EN couvertes par ce patron. Nous avons représenté (figure 11.3) ce taux de couverture en fonction du taux de réussite pour observer la couverture de discrimination d'un modèle vis-à-vis d'une classe : elle désigne le nombre d'exemples concernés par un patron, du point de vue de son taux de réussite.

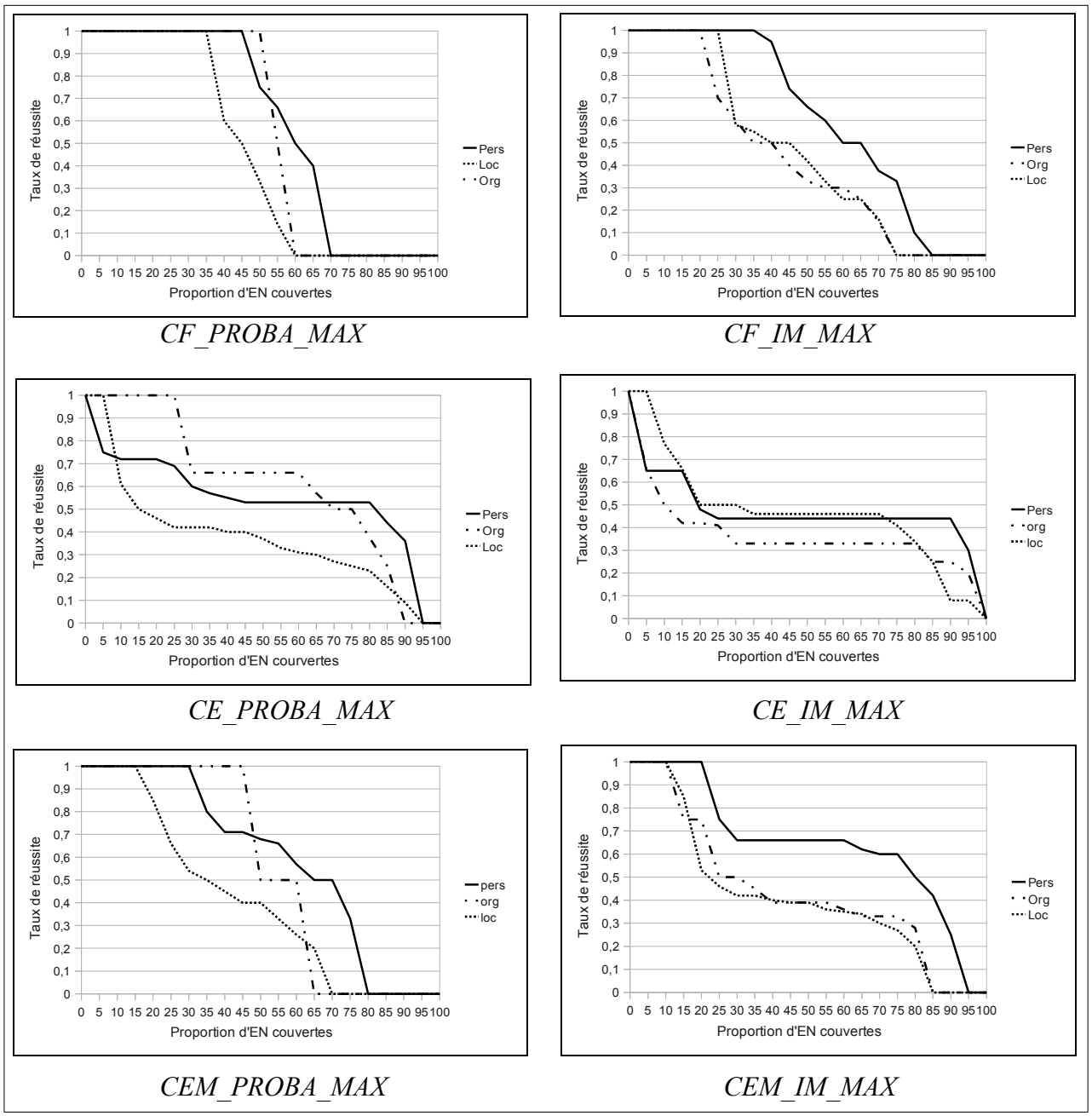


FIG 11.3 – Courbes de couverture de discrimination des 6 modèles MAX

Ces courbes nous indiquent le nombre d'occurrences d'EN couvertes (en pourcentage) en fonction de la capacité discriminante des patrons (peu importe leur nombre pour l'instant). On peut distinguer trois cas de figure majeurs :

- *le nombre d'occurrences couvertes par des bons patrons* (taux de 1) : par exemple, lorsque le modèle CF_PROBA_MAX génère des patrons pour la catégorie Organisation, 55% des occurrences sont couvertes par des bons patrons.
- *le nombre d'occurrences couvertes par des mauvais patrons* (taux de 0) : par exemple, lorsque le modèle CF_PROBA_MAX génère des patrons pour la catégorie Lieu ou Organisation, 40% des occurrences sont couvertes par des mauvais patrons.
- *le nombre d'occurrences couvertes par des patrons ambigus* (taux entre 0 et 1) : par exemple, lorsque le modèle CE_IM_MAX génère des patrons pour la catégorie Organisation ou Personne, 100% des occurrences sont couvertes par des patrons ambigus. Cette notion d'ambiguïté est à relativiser car le choix des patrons au sein des segments a également une influence, étant donné qu'un patron est associé au contexte duquel il est extrait. Le calcul s'effectuant sur l'intégralité du segment, la présence/absence d'éléments pertinents/non pertinents dans ce segment, influence indirectement le choix de ce patron.

On remarquera que les modèles IM extraient généralement moins de bons patrons et moins de mauvais patrons que les modèles PROBA, autrement dit, que leurs patrons sont généralement plus ambigus. On constate également que le Modèle CE est celui qui permet d'extraire le plus de patrons ambigus (entre 65% et 100% selon les cas). 25% des patrons de ce modèle sont néanmoins bons lorsqu'il est appliqué à la catégorie Organisation et combiné à la mesure PROBA.

Si on ne sélectionne que les bons patrons, on peut espérer corriger les erreurs du modèle de référence, et également juger de la pertinence des patrons correcteurs. Autrement dit, on pourra distinguer des patrons qui confirment les décisions du système Rnc et ceux qui les contredisent. Le nombre de patrons « infaillibles » par modèle et catégorie est indiqué dans le tableau (11.7).

Modèles	Nombre de Patrons				Nombre D'EN
	Lieu	Org	Pers	Total	
CF_PROBA_MAX	398	171	446	1015	1210
CEM_PROBA_MAX	243	112	297	652	789
CF_IM_MAX	202	74	263	539	674
CEM_IM_MAX	71	76	122	269	325
CE_PROBA_MAX	58	11	37	106	137
CE_IM_MAX	23	10	16	49	71

Tableau 11.5 – Nombre de bons patrons par modèle et classe d'EN

Le modèle CF_PROBA_MAX est celui qui produit le plus grand nombre de bons patrons (1015) lui permettant de couvrir correctement la moitié des EN du corpus de test. Le modèle CE produit peu de bons patrons mais le ratio entre le nombre de patrons et le nombre d'occurrences couvertes est plus élevé : ses patrons sont plus génériques. Néanmoins l'existence de bons patrons peut simplement s'expliquer par le fait qu'ils utilisent les mêmes indices que le système de référence –ce qui se traduirait par un taux de correction nul.

Les résultats présentés dans le tableau (11.8) indiquent les performances obtenues lorsque le système de référence détecte correctement une EN (R) et que les bons patrons sont appliqués. Les

résultats sont organisés en fonction de l'ajout de patrons issus de l'un des trois niveaux de représentation, du score d'association, ou tous modèles réunis. Lorsqu'aucun patron n'est identifié pour une instance d'EN donnée, le choix se porte sur la catégorie choisie par le système Rnc (R).

Catégorie	Modèle	Précision	Rappel	F-mesure	# Corrigés
LIEU	R+TOUS	0,767	0,945	0,847	91
	R+CF	0,754	0,930	0,833	78
	R+PROBA	0,726	0,939	0,819	86
	R+CE	0,727	0,910	0,808	61
	R+IM	0,739	0,886	0,806	41
	R+CEM	0,724	0,899	0,802	52
	R	0,670	0,836	0,744	*NA*
ORGANISATION	R+TOUS	0,952	0,686	0,797	121
	R+CF	0,941	0,672	0,784	107
	R+IM	0,917	0,666	0,772	101
	R+CE	0,925	0,624	0,745	61
	R+CEM	0,918	0,623	0,742	59
	R+PROBA	0,940	0,606	0,737	45
	R	0,866	0,561	0,681	*NA*
PERSONNE	R+TOUS	0,924	0,978	0,950	31
	R+CF	0,916	0,975	0,944	27
	R+PROBA	0,909	0,977	0,942	30
	R+CE	0,898	0,970	0,932	21
	R+IM	0,897	0,967	0,930	18
	R+CEM	0,894	0,969	0,930	20
	R	0,861	0,951	0,904	*NA*

Tableau 11.6 – Potentiel de correction du système de référence.

Globalement, la prise en compte de tous les patrons de correction permet d'améliorer les f-mesures de 5% pour les Personnes, et de 10% pour les Lieux et les Organisations. Le niveau de représentation qui permet de corriger le plus d'instances est le niveau CF, en partie du fait qu'il génère un plus grand nombre de patrons, multipliant ainsi les possibilités de désambiguïsation. Le taux de correction par modèle ne dépend pourtant pas uniquement du nombre de patrons extraits : les niveaux CE et CEM ont un taux de correction relativement équivalent alors que le niveau CEM génère un plus grand nombre de patrons. Ceci s'explique simplement par le fait qu'une large part des patrons corrects de ce dernier vient confirmer le modèle de référence. La combinaison de tous les modèles permet de corriger 66% d'erreurs pour les Lieux, 55% pour les Personnes et 28% pour les Organisations.

Ces résultats semblent corroborer le lien entre la performance d'un modèle et sa capacité de correction : les modèles ayant permis de corriger un grand nombre d'organisations sont basés sur le score IM, alors que le score PROBA contribue à générer plus de patrons de correction pour les lieux et les personnes. Nous avions constaté des tendances similaires sur les diagrammes de la figure (11.2). Cela nous pousserait à considérer les personnes et les lieux comme similaires (PROBA) et différents des organisations (IM). Pour interpréter ce phénomène, il faut garder en mémoire, d'une part, que les patrons de correction sont des patrons qui infirment les décisions du système Rnc, et d'autre part, que l'IM a la propriété de ne pas dépendre directement de la fréquence pour extraire les patrons, à l'inverse du score PROBA. Par conséquent, les patrons IM auraient plus de chances d'être distants (les performances croissent également en fonction de la taille du segment) ou rares, à l'inverse des patrons PROBA qui seraient plus systématiques et indifférents à la distance. Une telle corrélation paraît difficile à établir, mais elle signifierait que certains éléments du contexte des organisations sont foncièrement différents de ceux des deux autres classes.

11.2.3. Analyse des patrons correcteurs

Le point que nous souhaitons éclaircir à présent est de savoir si les patrons correcteurs réalisent véritablement des relations sémantiques avec les EN ou si leur association est fortuite. Dans ce qui suit, nous ne présenterons qu'une sélection concise des patrons correcteurs, en tentant d'identifier les caractéristiques majeures des modèles. Nous ne présenterons pas le modèle CEM, étant donné qu'une majeure partie de ses patrons sont couverts par les modèle CE (pour la catégorie d'entité-R) et CF (pour la catégorie de forme). Le tableau (11.9) montre quelques exemples de patrons correcteurs obtenus pour le modèle CE en fonction des divergences de décision entre le modèle (CAT(MODEL)) et le système de référence (CAT(R)), lorsqu'il y a lieu.

MODELE	Patrons	CAT(MODEL)	CAT(R)
PROBA	*GNP_à/_val_unitDist*	LOC	PERS
	GNP_de/_Tpopulation	LOC	ORG
	GNP_des/_pers_fonct	ORG	LOC
	_acro_div	ORG	PERS
	GNP_du/_titre	PERS	LOC
	GNP_durant/_subs	PERS	ORG
IM	*_LLoc*	LOC	PERS
	GNP_lors d'/_subs	LOC	ORG
	_range_objet_complet	ORG	LOC
	GNP_par/_subs	ORG	PERS

Tableau 11.7 – Exemples de patrons correcteurs pour le modèle CE

Le tableau s'interprète comme suit : le patron *GNP_à/_val_unit_dist* désignant un groupe prépositionnel introduit par *à* dont la tête est une entité-R de valeur de distance est associé plus significativement à la catégorie *Lieu*, ce qui a permis de corriger (au moins) une instance d'EN annotée comme *Personne* par le système Ritel-nca (CAT(R)). Le segment correspondant est illustré en (257) (patron en bleu, EN en rouge).

(257)à un ou deux kilomètres de Ras Shamra

Nous avons projeté ce patron sur l'intégralité du corpus de test pour évaluer son intérêt : il permet de discriminer des noms d'îles à partir de noms ambigus désignant des personnes (258 et 259) ou des sièges d'organisations (260 et 261).

(258)à trente kilomètres au nord de Saint - Barthélemy vit son île cousine
(259)un groupe de 50 dauphins a été repéré à une cinquantaine de mètres de la plage de Maria
(260)se trouve à une dizaine de kilomètres de Toyota City
(261)à une centaine de mètres de la Fondation Pierre Bergé - Yves Saint Laurent

Ce patron capte donc une véritable relation entre la distance et le lieu : la présence d'un tel patron dans un segment indiquera avec une forte certitude que si une EN est également présente, elle sera de type Lieu.

Si l'on s'intéresse à un patron extrait par IM, comme *_LLoc*, qui désigne un type de lieu générique (correspondant aux formes *pays*, *ville*, etc.), on trouve l'exemple (262).

(262)le pays de Bush

Conformément aux conventions d'annotation, on annote le Npr en fonction du déclencheur *pays* ; il s'agit donc d'un lieu. Effectivement, lorsque l'on limite la taille des segments contenant le déclencheur à 2, l'EN est très souvent un lieu dans le corpus de test (263). Néanmoins on trouve des occurrences où il est erroné lorsqu'il est en seconde position ou que l'EN incluse dans un GNP n'est pas introduite par la préposition *de* ; dans de tels cas, il est employé comme verbe (264).

(263)hôtel Drouot, cité Aubry, avenue du Président – Wilson, rue d' Isly à Alger, avenue des Grésillons
(264)Claude Pinganaud adresse, cité par l' agence d' information Lenta

La portée de ce patron nécessite d'être restreinte à un contexte local dans des segments plus longs comme on peut l'observer dans l'exemple (265).

(265)Derrida pouvait descendre dans la rue accompagné de François Châtelet

Si l'on se concentre à présent sur le modèle CF pour la classe Personne, on trouve des prénoms comme *Véronique* (266) et des formes verbales comme *réplique* (267) ou *observe* (268).

(266)Isham et Véronique
(267)réplique Eiffel
(268)observe Maud

Les formes verbales employées comme patron pour ce qui concerne la classe Organisation sont plus souvent au pluriel : *reprochent* (267), *savent* (268), *semblent* (269) sont des exemples.

(269)les juges reprochent à la Communauté urbaine de Strasbourg
(270)connaissent l' histoire de l' Espagne savent
(271)les Etats - Unis semblent vouloir reproduire dans le domaine naval le modèle du programme américain Joint Strike Fighter

On trouve également des noms de fonctions (*PDG*, *président, stratège*), des noms de types d'organisation (*parti*, *éditions*, *filiale*, etc.) des noms propres d'organisation (*El Mundo*, *Dow Jones*, etc.) ou encore des valeurs monétaires (*7 millions d'euros*, etc.). Il y a cependant des exemples plus étranges comme *1940*, *ceux*, ou *actuellement*. Avant de nous prononcer sur leur pertinence, observons les segments desquels ils ont été extraits.

(272)sa trilogie sur l' Algérie en pleine fièvre nationaliste à la fin des années 1940 a été adaptée en feuilleton
(273)ceux d' Espace Marx
(274)présente actuellement le Louvre

Les trois exemples sont des cas d'ambiguïté métonymique Lieu/Organisation. Dans l'exemple (273), *ceux* désigne les locaux appartenant à l'association *Espace Marx*. Le contexte de l'exemple (274) est une exposition que présente l'organisation qui gère le Louvre. Rappelons que pour sélectionner ces patrons, les modèles comparent les valeurs de tous les éléments au sein du segment pour sélectionner celui ayant la plus forte valeur et lui assigner ainsi une catégorie. Dans le cas de *ceux*, qui est le seul élément, cela signifie que son score est plus important pour la catégorie *Organisation* ; pour l'exemple (274) cela signifie que *actuellement* a un meilleur score pour la

catégorie *Organisation* et que ce score est plus élevé que les scores possibles pour l'élément *présente*. Enfin, pour l'exemple (272), cela signifie qu'aucun élément du segment n'a un score plus important que celui de *1940* pour la catégorie *Organisation*. Certes, la décision du modèle résulte d'un calcul complexe (que nous avons tâché de garder simple) qui pourrait trouver une interprétation, mais ces patrons ne sont pas interprétables tels quels et ils nous permettent encore moins de construire des règles sensées. Ces patrons « fortuits » nous ramènent aux résultats que nous avions obtenus en (3.2) lorsque nous avions analysé les collocations du verbe *abandonner*.

Pour la catégorie Lieu du modèle CF, on trouve des noms propres de lieux comme *au Pays basque* (275), des participes associés à la notion de localisation, comme *basée* (276), *décollant* (277) ou *exposée* (278), ou encore des locutions prépositionnelles non rattachées comme *au fond* (279). Certains patrons fortuits ont également permis de classer correctement des noms de lieux, comme *je* (280).

(275)cinq personnes au Pays basque et dans le Léon
(276)basée à Saint - Apollinaire
(277)quelques charters décollant directement de Lampedusa vers Le Caire ou Tripoli avaient déjà été organisés dans le passé
(278)exposée en septembre dernier à la galerie Colette
(279)l' homme labourait au fond de la baie de Minet El - Beida
(280)et maintenant je suis à Yale

Pour conclure, les patrons correcteurs, s'ils permettent d'améliorer les résultats, doivent être maniés avec précaution. Bien qu'ils permettent de corriger des EN, ils ne sont pas toujours pertinents sur le plan sémantique. Ils offrent tout de même d'intéressantes associations entre des entités-R de type *fonction* et *Organisation* ou encore entre *Lieu* et *distance*. Bien évidemment nous n'avons analysé que les patrons ayant permis de corriger les EN fournies par le système Ritel-nca, ce qui laisse du champ à l'analyse des bons patrons, ayant confirmé des décisions du système de référence dans le cadre de l'évaluation, mais pouvant être réemployés dans d'autres contextes. L'analyse des mauvais patrons combinée avec les erreurs du système Rnc peut aussi révéler d'intéressantes corrélations.

11.3. Bilan

Ce volet présente une seconde application du modèle de segmentation discursive, l'analyse intra-segment. Plutôt que d'établir un contexte de recherche arbitraire de *n* mots autour d'un mot-cible, ce modèle permet de définir une telle fenêtre sur des critères linguistiques, les segments. L'analyseur présenté s'appuie sur trois niveaux d'information linguistique fournis par l'analyseur Ritel-nca et le chunking pour définir des patrons sémantiques qui désambiguïsent et corrigent une Entité Nommée. Le système proposé présente en effet l'originalité de ne s'appuyer sur d'autres informations que les résultats des systèmes antérieurs, pour effectuer sa propre classification. Nous faisons l'hypothèse que les patrons qui y parviendront seront liés par une relation sémantique à l'Entité Nommée.

Les faibles performances obtenues peuvent être interprétées comme les résultats d'une utilisation de mesures de scores simples (Probabilité et Information Mutuelle), mais elles peuvent aussi être imputées au fait que ses entrées comportent des erreurs qui se répercutent sur les décisions du système. Le système n'est pas à proprement parler supervisé (voir [Petasis et al., 2001] pour une perspective similaire) car nous avons souhaité analyser les patrons apparaissant localement dans les segments et évaluer leur pertinence linguistique. Nous avons mesuré le potentiel de correction du système dans le cas où les meilleurs patrons étaient sélectionnés. Si un grand nombre confirme les décisions du système de référence, d'autres permettent de le corriger, ce qui améliorerait les performances sur les trois classes d'EN. Pour s'en assurer, il nous faudrait mener une nouvelle évaluation en testant ces patrons acquis automatiquement à partir de corpus.

Nous avons proposé une évaluation manuelle en projetant les patrons sur l'intégralité du corpus de test. Si certaines associations sont fortuites (certains adverbes par exemple), le système a permis d'extraire des relations sémantiques pertinentes (lieu et distance par exemple). Une autre possibilité consisterait à restreindre les entrées du système EnCor par des exemples que l'on sait corrects, en filtrant les exemples avec une liste d'EN pour que le processus reste automatique. Une telle méthode serait assez similaire aux travaux de E. Riloff et Jones [Riloff & Jones, 1999], sauf que ces derniers définissent leurs patrons sur la base de relation de dépendance.

Un des inconvénients de cette méthode est de ne pas pouvoir traiter les segments simple (de taille 1), qui ne sont pas évalués. Cette évaluation appelle donc d'autres expériences prenant en compte les relations inter-segment, ainsi que l'impact de la grammaire de segment sur l'extraction de patrons, afin de valider l'intérêt de la segmentation pour la classification des EN.

Enfin, la caractérisation des conventions d'annotation constituent l'étape primordiale de l'évaluation des systèmes. De notre point de vue, elles doivent être réalistes, mais aussi fournir des critères fiables pour l'annotation. Les critères linguistiques permettent notamment de préciser la définition des catégories en fonction de leur usage en corpus. Pour compléter ces critères, nous avons mis en perspective ces conventions avec le cadre de notre domaine d'application, les SQR : en évaluant les questions auquel un exemple était susceptible de répondre, cela nous permet de prendre une décision de classification pour une majorité des cas.

CONCLUSION

Le propos principal de ce travail a été de montrer l'importance du rôle joué par le contexte dans l'extraction de relation sémantique. Par contexte, nous entendons d'une part, l'environnement linguistique d'un mot et d'autre part, le genre du corpus d'étude. L'environnement linguistique peut être caractérisé sur plusieurs dimensions ; celles que nous avons étudiées sont la forme (collocation), la syntaxe (chunking et dépendances) et la sémantique (catégories ontologiques). Ces dimensions permettent d'affiner les relations sémantiques entre mots, mais se révèlent insuffisantes lorsque leur définition ne prend pas en compte le corpus. Nous avons montré que la caractéristique du genre, comme le conte ou l'article de presse, avait un impact sur les relations sémantiques. En effet, selon le genre des textes, la distribution syntaxique des catégories sémantiques, ce que nous avons décrit comme l'alternance de types sémantiques, change. Cette variation ne peut être observée par un modèle linguistique qui ne prend pas en compte le genre. Pour concevoir un système d'extraction de relation sémantique opérationnel qui puisse être utilisé dans des contextes différents, la variable du genre ne peut donc être ignorée.

En outre, l'Extraction d'Information (EI) nous a appris que le genre n'est pas l'unique variable conditionnant les relations sémantiques en corpus et que la prise en compte du domaine est probablement tout aussi importante. En effet, pour fonctionner, l'extraction d'information suppose la présence d'éléments des cadres (comme les transactions financières ou les attaques terroristes) dans les collections de textes employés. Le cadre implique un domaine spécifique auquel appartient un texte ; or, ce dernier limite le nombre de relations sémantiques qu'il sera possible d'extraire dans le texte. Les Entités Nommées (EN), que nous considérons comme des noms propres sémantiquement catégorisés, font partie des informations primaires qui sont mises en relation dans les cadres d'EI.

La tâche de reconnaissance et de classification des EN (RCEN) a été identifiée comme une technologie générique qui puisse être employée à tout domaine et appliquée à tout type de texte. Cet objectif de généricité aura eu pour conséquence d'affranchir les EN des variables du domaine et du genre. Par conséquent, en les réutilisant dans d'autres tâches, il est nécessaire de s'interroger sur leur adéquation. Les systèmes de question-réponse (SQR) qui s'appuient sur la classification des systèmes de RCEN pour extraire les réponses ont constitué notre contexte d'application.

Notre étude des conventions d'évaluation des systèmes de RCEN a mis en évidence les contradictions auxquelles ils se heurtent, étant donné que les catégories sémantiques changent en contexte. Les recherches récentes sur la métonymie des EN pointent en effet la faiblesse d'un système qui ne prend pas en compte cette variabilité. Il a été démontré qu'en s'appuyant sur des indices syntaxico-sémantiques, il était possible d'identifier et d'expliquer ces glissements de sens que produit la métonymie.

Cependant, les évaluations officielles n'ont été menées que sur des corpus de presse. Si nos analyses sur l'impact du genre sont justes, la conception de systèmes de RCEN génériques devra mettre en corrélation les alternances sémantiques en fonction du genre de texte.

Nous avons proposé de (re-)situer la RCEN dans le cadre de l'EI parce que les relations sémantiques permettent de mieux appréhender cette tâche. Ce « retour au sources » nous a permis de souligner le lien qui existe entre les systèmes d'EI et les SQR : l'extraction de réponse à des questions factoïdes peut être liée à une relation sémantique d'un cadre d'EI en corpus.

En plus de ces problèmes de modélisation sémantique, l'extraction de relations sémantiques fait face à la complexité des structures discursives de surface auxquelles elles sont mêlées. En effet, une relation sémantique n'associe pas toujours des mots qui sont textuellement adjacents : ces derniers peuvent apparaître à une distance variable l'un de l'autre. Le système doit par conséquent être suffisamment élastique pour autoriser des substitutions dans l'ordre de succession des éléments, mais également détecter des relations à une distance qu'on ne peut prédéfinir. Or, les systèmes symboliques auxquels nous nous sommes intéressé s'appuient sur des grammaires dont la rigidité entraîne la multiplication des règles. Par ailleurs, la détection de relation à longue distance est un problème fondamental qui constitue une limite actuelle des systèmes de TAL en général et des systèmes d'extraction de relation en particulier.

Pour répondre à ces problèmes, nous avons proposé un modèle de segmentation discursif de surface qui consiste à fragmenter le texte sur des critères linguistiques. Les frontières expérimentées pour la délimitation de ces segments sont la ponctuation, les conjonctions et connecteurs discursifs. En intégrant un segmenteur à notre système, nous avons pu raisonner au niveau discursif et identifier des structures discursives comme les insertions, les listes ou encore les parenthétiques, à partir de grammaires de segments. Comme nous l'avons vu, ces structures discursives renferment des relations sémantiques, mais leur identification permet également de réduire la distance entre des éléments associés par une relation sémantique.

Le système d'extraction de relations sémantiques proposé s'appuie sur les informations sémantiques fournies par l'analyseur linguistique Ritel-nca (Rnc), les informations syntaxiques de l'analyseur en chunk LoRit et les informations discursives du segmenteur. Il s'agit d'un système hybride. À partir d'un formalisme commun, il permet à la fois de concevoir des règles symboliques et d'en extraire automatiquement à partir de corpus. L'extraction automatique s'appuie sur les données obtenues sur un corpus par les systèmes en amont et sur une modélisation nivelée des informations au sein des segments. Les patrons sémantiques obtenus sont associés à un score qui traduit la dépendance entre deux éléments potentiellement en relation.

Deux évaluations de ce système ont été menées : la détection de relations à longue distance et l'adaptation d'un système de RCEN.

La détection de relation à longue distance est un problème complexe, qui repose à la fois sur la détection de relations syntaxiques et la catégorisation sémantique des entités syntaxiquement liées. Pour évaluer notre système, nous avons donc dû concevoir une grammaire syntaxique. L'évaluation a porté sur l'augmentation des relations sujet obtenues par cette grammaire, lorsque la grammaire de segments était utilisée. Bien que l'évaluation n'ait porté que sur un seul verbe, elle a clairement montré l'influence des structures discursives sur l'extraction de relations, ainsi que l'apport d'un traitement discursif pour l'augmentation de ces relations.

L'adaptation d'un système de RCEN a consisté à définir une tâche en adéquation au contexte d'application de l'analyseur linguistique Rnc : la recherche d'information précise dans un contexte de dialogue. Nous avons déterminé des conventions d'annotation en fonction des types de question auxquelles une phrase pourrait répondre. L'évaluation a montré que les performances du système de référence Rnc étaient dégradées et que le correcteur pouvait corriger ces erreurs.

Plus généralement, nous espérons avoir confirmé l'apport de la linguistique dans la caractérisation des problèmes sémantiques auxquels sont confrontés les systèmes de TAL, ainsi que son intérêt dans le cadre de leur opérationalisation. Néanmoins, les modèles linguistiques nécessitent d'être réinterprétés dans le contexte d'application particulier pour que puisse véritablement être identifié leur intérêt. Par ailleurs, il convient de noter que les limites auxquelles se heurtent les systèmes correspondent aussi à des problématiques marginales en linguistique : nous avons isolé les noms propres, la ponctuation, les citations. Le TAL permet donc de mettre en lumière les carences éventuelles des modèles linguistiques lorsque l'on cherche à les appliquer de manière systématique. Si de tels problèmes émergent, c'est parce que le TAL soulève des questions fondamentales sur l'interaction entre texte et référence, et sur ce que Rastier nomme la composante tactique du texte.

Le lien entre texte et référence soulève des problèmes de catégorisation sémantique du langage : comment le langage permet-il de parler de la réalité et de quelle réalité s'agit-il ? Nous avons montré que les ressources de connaissances linguistiques, comme les lexiques syntaxiques, s'appuyaient sur des considérations ontologiques, qui calquent leur classification sur la réalité L'exploitation de ces ressources pour des corpus relevant d'un genre spécifique, montre que ces catégories sont trop générales et qu'elles ne permettent pas de prédire correctement les relations sémantiques. À travers l'analyse des contes, nous avons mis en évidence que les écarts entre classification ontologique et corpus étaient quantitativement trop importants pour être ignorés. Les ontologies isolent en effet très tôt les êtres imaginaires ou fictifs qui pourtant se comportent comme des êtres humains dans les contes. Mais les êtres imaginaires ne sont pas une exception, les animaux, les objets, les végétaux montrent le même phénomène.

Si l'on veut pouvoir analyser les relations sémantiques dans des genres et des domaines différents, il faut disposer de connaissances spécifiques sur ces textes. Tout l'apport de la linguistique de corpus est de ne pas imposer des grilles d'interprétation aux textes, mais de faire émerger en priorité les relations qui s'établissent dans le texte avant de les interpréter. Par conséquent, il est crucial de poursuivre des analyses sémantiques de corpus pour mettre à jour plus finement les liens qui se tissent dans l'univers d'un texte. La linguistique de corpus joue fort probablement un rôle-clé entre la linguistique et le TAL. Par son approche quantitative, elle permet de faire émerger le sens des mots à partir de l'usage qui en est fait en corpus, soit sur des données attestées. Ce faisant, elle interroge les rapports entre théorie et instance, et analyse les interactions entre modèle linguistique et texte. Comme nous l'avons proposé, les analyses de corpus peuvent être enrichies en combinant les méthodes avec des ressources issues du TAL.

Une recherche qui mutualiserait à la fois TAL, Linguistique et Corpus permettrait par conséquent d'alimenter les réflexions dans les trois domaines et on peut l'espérer, de dépasser les limites auxquelles elles sont confrontées isolément. Nous pensons avoir contribué à cet objectif d'identifier des passerelles entre informatique et linguistique, pour ces disciplines qui, rappelons-le, partagent le même objet d'étude : le Langage.

BIBLIOGRAPHIE

ACE2002, 2002, *ACE 2002 Evaluation Plan*, p. 1-6.

ADAM Jean-Michel, 2001, « Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ? », *Langages*, vol. 35, no. 141, p. 10-27.

ANTOINE Jean-Yves, 2003, *Pour une ingénierie des langues plus linguistique*, Mémoire d'Habilitation à Diriger les Recherches, , Lorient/Vannes.

AÏT-MOKHTAR Salah, CHANOD Jean-Pierre and ROUX Claude, 2002, « Robustness beyond shallowness: incremental deep parsing », *Natural Language Engineering*, vol. 8, no. 3, p. 121–144.

BARNBROOK Geoff, 2002, *Defining language: a local grammar of definition sentences*, Amsterdam, John Benjamins Publishing Company.

BAYRAKTAR Murat, SAY Bilge and AKMAN Varol, 1998, « An Analysis of English Punctuation: The Special Case of Comma », *International Journal of Corpus Linguistics*, vol. 3, no. 1, p. 33-57.

DE BEAUGRANDE Robert, 1991, *Linguistic theory : the discourse of fundamental works*, London & New York, Longman.

BENAJIBA Yassine, ZITOUNI Imed, DIAB Mona T. and ROSSO Paolo, 2010, « Arabic Named Entity Recognition: Using Features Extracted from Noisy Data », *ACL*, 2010, p. 281-285.

BENNACEF Samir, BONNEAU-MAYNARD Hélène, GAUVAIN Jean-Luc, LAMEL Lori and MINKER Wolfgang, 1994, « A Spoken Language System for Information Retrieval », *International Conference on Spoken Language Processing (ICSLP)*, 1994, p. 1271–1274.

BERNARD Guillaume, 2011, *Réordonnancement d'hypothèses dans un système de Questions-Réponses*, Thèse de doctorat, Université Paris Sud (Limsi), Paris.

BIBER Douglas, 1991, *Variation across speech and writing*, Cambridge, Cambridge University Press.

—, 1993, « Representativeness in Corpus Design », *Literary and Linguistic Computing*, vol. 8, no. 4, p. 243 -257.

BIKEL Daniel M., SCHWARTZ Richard and WEISCHEDEL Ralph M., 1999, « An Algorithm that Learns What's in a Name », *Machine Learning*, vol. 34, no. 1-3, p. 211–231.

BLOOMFIELD Leonard, 1933, *Language*, Chicago, University Of Chicago Press.

BONAMI Olivier and GODARD Danièle, 2008, « On the Syntax of Direct Quotation in French », 15th International Conference on HPSG, 2008, Stanford, CSLI Publications, p. 358-377.

BORTHWICK Andrew E., 1999, *A maximum Entropy Approach to Named Entity Recognition*, Thèse de

doctorat, New York University, New York.

Bourigault Didier, 2002, « Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », TALN, 2002, Nancy, p. 75-84.

—, 2007, *Un analyseur syntaxique opérationnel : SYNTEX*, Mémoire d'Habilitation à Diriger les Recherches, Université Toulouse II-Le Mirail, Toulouse.

Brun Caroline, Ehrmann Maud and Jacquet Guillaume, 2007, « XRCE-M: a hybrid system for named entity metonymy resolution », *4th International Workshop on Semantic Evaluations*, 2007, Stroudsburg, PA, USA, Association for Computational Linguistics (SemEval '07), p. 488–491.

Bréal Michel, 1897, *Essai de sémantique : science des significations*, Paris, Hachette.

Bunescu Razvan C. and Pasca Marius, 2006, « Using Encyclopedic Knowledge for Named entity Disambiguation », *EACL*, 2006.

Burton Richard R., 1977, « Semantic grammar: an engineering technique for constructing natural language understanding systems », *SIGART Bulletin*, no. 61, p. 26–26.

Bühler Karl, 2009, *Théorie du langage : la fonction représentationnelle*, Marseille, Agone.

Cadiot Pierre and Habert Benoît eds., 1997, *Aux sources de la polysémie nominale* (Langue Française), n° 113.

Charolles Michel, 2002, *La référence et les expressions référentielles en français*, Paris, Ophrys.

Chinchor Nancy, 1998, « MUC-7 Named Entity Task Definition », *the Seventh Message Understanding Conference MUC-7*, MUC-7, 1998, Fairfax, VA.

Chiticariu Laura, Krishnamurthy Rajasekar, Li Yunyao, Reiss Frederick and Vaithyanathan Shivakumar, 2010, « Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks », *EMNLP*, 2010, p. 1002-1012.

Chomsky Noam, 1965, *Aspects of the theory of syntax*, Cambridge, MA, MIT Press.

—, 1970, « Remarks on nominalization », *Readings in English transformational grammar*, R. A. Jacobs and P. S. Rosenbaum eds., Waltham, MA, Blaisdell, p. 184-221.

Church Kenneth W. and Hanks Patrick W., 1990, « Word Association Norms, Mutual Information, and Lexicography », *Computational Linguistics*, vol. 16, no. 1, p. 22-29.

Collins Michael J., 1996, « A new statistical parser based on bigram lexical dependencies », *34th annual meeting on Association for Computational Linguistics*, 1996, Stroudsburg, PA, USA, p. 184–191.

Collins Michael J. and Singer Yoram, 1999, « Unsupervised Models for Named Entity Classification », *the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, p. 100–110.

Cruse Alan, 1986, *Lexical semantics*, Cambridge, Cambridge University Press.

—, 1996, « La signification des noms propres de pays en anglais », *Les mots de la nation*, S. Rémi-Giraud and P. Rétat eds., Lyon, Presses Universitaires de Lyon, p. 93-102.

Cucchiarelli Alessandro and Velardi Paola, 2001, « Unsupervised named entity recognition using syntactic and semantic contextual evidence », *Computational Linguistics*, vol. 27, no. 1, p. 123–131.

Culioli Antoine, 1990, *Pour une linguistique de l'énonciation: opérations et représentations*, Paris, Ophrys.

Doddington George, Mitchell Alexis, Przybocki Mark, Ramshaw Lance, Strassel Stephanie and Weischedel Ralph, 2004, « The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation », *LREC*, 2004, p. 837–840.

Donnat Olivier, 2009, *Les pratiques culturelles des Français à l'ère numérique : enquête 2008*, Paris, La Découverte/Ministère de la Culture et de la Communication.

Dowty David, 1991, « Thematic Proto-Roles and Argument Selection », *Language*, vol. 67, no. 3, p. 547-619.

Dubreil, 2008, *La dimension argumentative des collocations textuelles en corpus électronique spécialisé au domaine du TAL(N)*, Thèse de doctorat, Université de Nantes, Nantes.

Dunning Ted, 1993, « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics*, vol. 19, no. 1, p. 61–74.

ESTER2, 2007, *Conventions d'annotation des entités nommées ESTER 2*, AFCP, DGA/CEP, ELDA, p. 1-22.

Ehrmann Maud, 2008, *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Thèse de doctorat, Université Paris 7, Paris.

Engler Rudolf, 1973, « Rôle et place d'une sémantique dans une linguistique saussurienne », *Cahiers Ferdinand de Saussure*, no. 28, p. 35-52.

Fabre Cécile and Frérot Cécile, 2002, « Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus », *TALN*, 2002, Nancy.

Fillmore Charles J., 1977, « Scenes-and-frames semantics », *Linguistic Structures Processing*, A. Zampolli ed., Amsterdam, North Holland, p. 55-81.

—, 1982, « Frame Semantics », *Linguistics in the Morning Calm*, SICOL 1981, 1982, Séoul, The Linguistic Society of Korea, p. 111-137.

—, 2003, *Form and meaning in language*, Stanford (Ca), CSLI Publications.

Firth John R., 1957, *Papers in linguistics, 1934-1951.*, London/New York, Oxford University Press.

—, 1968, *Selected papers of J.R. Firth, 1952-1959*, Bloomington/London, Indiana University Press.

Francis Nelson W., 1992, « Language Corpora B.C. », *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, J. Svartvik ed., Berlin/New York, Mouton de Gruyter (Trends in Linguistics. Studies and Monographs), p. 17-32.

Frege Gottlob, 1971, *Ecrits logiques et philosophiques*, Paris, Editions du Seuil.

Friburger Nathalie, 2002, *Reconnaissance automatique des noms propres. Application à la classification automatique des textes journalistiques*, Thèse de doctorat, Université François Rabelais, Tours.

Gaizauskas Robert J. and Wilks Yorick, 1998, « Information Extraction: Beyond Document Retrieval », *Journal of Documentation*, vol. 54, no. 1, p. 70-105.

Galibert Olivier, 2009, *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*, Thèse de doctorat, Université Paris Sud (Limsi), Paris.

Galibert Olivier, Quintard Ludovic, Rosset Sophie, Zweigenbaum Pierre, Nedellec Claire, Aubin Sophie, Gillard Laurent, Raysz Jean-Pierre, Pois Delphine, Tannier Xavier, Deléger Louise and Laurent Dominique, 2010, « Named and Specific Entity Detection in Varied Data: The Quæro Named Entity Baseline Evaluation », *LREC*, 2010, Malte, p. 3453-3458.

Galliano Sylvain, Gravier Guillaume and Chaubard Laura, 2009, « The ester 2 evaluation campaign for the rich transcription of French radio broadcasts », *INTERSPEECH*, 2009, p. 2583-2586.

Garde Paul, 1985, « Dualité de la relation syntaxique : relation dépendantielle et relation référentielle », *Les relations syntaxiques*, Travaux du Cercle Linguistique d'Aix-en-Provence, p. 1-25.

Garera Nikesh and Yarowsky David, 2009, « Structural, Transitive and Latent Models for Biographic Fact Extraction », *EACL*, 2009, p. 300-308.

Gary-Prieur Marie-Noëlle, 1994, *Grammaire du nom propre*, Paris, Presses Universitaires de France.

Gavaldà Marsal, 2000, *Growing semantic grammars*, Phd Thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Geierhos Michaela, Blanc Olivier and Bsiri Sandra, 2009, « RELAX Extraction de relations sémantiques dans les contextes biographiques », *TAL*, vol. 49, no. 1, p. 167-190.

Gildea Daniel and Jurafsky Daniel, 2002, « Automatic Labeling of Semantic Roles », *Computational Linguistics*, vol. 28, no. 3, p. 245-288.

Givón Talmy, 1985, « Iconicity, Isomorphism and Non-arbitrary Coding in Syntax », *Iconicity in Syntax*, J. Haiman ed., Amsterdam, John Benjamins, p. 187–219.

Grefenstette Gregory, 1994, *Explorations in automatic thesaurus discovery*, Norwell (Ma), Springer.

Grishman Ralph, 2010, « The Impact of Task and Corpus on Event Extraction Systems », *LREC*, 2010.

Grishman Ralph and Sundheim Beth M., 1995, « Design of the MUC-6 evaluation », *MUC*, 1995, p. 1-11.

Gross Maurice, 1975, *Méthodes en syntaxe: régime des constructions complétives*, Paris, Hermann.

Grouin Cyril, Deléger Louise, Cartoni Bruno, Rosset Sophie and Zweigenbaum Pierre, 2011, « Accès au contenu sémantique en langage de spécialité : extraction des prescriptions et concepts médicaux », *TALN*, 2011, p. 109–120.

Guillaume Gustave, 1973, *Principes de linguistique théorique de Gustave Guillaume*, Québec/Paris, Presses de l'Université Laval et Klincksieck.

Halliday Michael A. K., 1966, « Lexis as a linguistic level », *In Memory of J.R. Firth*, C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins eds., London, Longman, p. 148-163.

—, 1994, *An Introduction to Functional Grammar*, 2nd ed., London, Arnold.

Halliday Michael A. K. and Hasan Ruqayia, 1976, *Cohesion in English*, London, Longman.

Hanks Patrick W., 1997, « Lexical Sets: Relevance and Probability », *Translation and Meaning. Part 4*, Lodz Session of the 1st International Maastricht-Lodz Duo Colloquium on "Translation and Meaning," 1997, Maastricht.

—, 2008, « Lexical Patterns: From Hornby to Hunston and Beyond », Euralex, 2008, Barcelone, p. 89-129.

Hanks Patrick W. and Ježek Elisabetta, 2008, « Shimmering lexical sets », *Euralex*, 2008, Barcelone, p. 391-402.

Hanks Patrick W., Pala Karel and Rychly Pavel, 2007, « Towards an empirically well-founded ontology for NLP », *4th International Workshop on Generative Approaches to the Lexicon*, Paris, 2007.

Harris Zellig, 1951, *Methods in Structural Linguistics*, Chicago, University of Chicago Press.

Hearst Marti A., 1992, « Automatic acquisition of hyponyms from large text corpora », *14th conference on Computational linguistics - Volume 2*, 1992, Stroudsburg, PA, USA, Association for Computational Linguistics (COLING '92), p. 539–545.

Hoey Michael, 2005, *Lexical priming : a new theory of words and language*, London/New York, Routledge.

Hunston Susan, 2003, « Frame, phrase or function: a comparison of frame semantics and local grammars », *Corpus Linguistics Conference*, 2003, Lancaster, p. 342-358.

—, 2011, *Corpus approaches to evaluation: phraseology and evaluative language*, London, Routledge.

Hunston Susan and Francis Gill, 2000, *Pattern grammar: a corpus-driven approach to the lexical grammar of English*, Amsterdam, Benjamins.

Hunston Susan and Sinclair John McH., 2000, « A Local Grammar of Evaluation », *Evaluation in text: authorial stance and the construction of discourse*, G. Thompson and S. Hunston eds., Oxford, Oxford University Press, p. 75-100.

Isozaki Hideki and Kazawa Hideto, 2002, « Efficient support vector classifiers for named entity recognition », *19th international conference on Computational linguistics - Volume 1*, 2002, Stroudsburg, PA, USA, Association for Computational Linguistics (COLING '02), p. 1–7.

Jackendoff Ray S., 1985, *Semantics and cognition*, Cambridge (Ma), MIT Press.

Jakobson Roman, 1963, *Essais de linguistique générale. 1. Les fondations du langage*, Paris, Les Editions de Minuit.

Jonasson Kerstin, 1994, *Le nom propre: constructions et interprétations*, Louvain-la-Neuve, Duculot.

Kachru Braj B., 1980, « "Socially realistic linguistics": the Firthian tradition », *Studies in the Linguistic Sciences*, vol. 10, no. 1, p. 85-111.

Katz Jerrold J. and Fodor Jerry A., 1963, « The Structure of a Semantic Theory », *Language*, vol. 39, , p. 170–210.

Kilgarriff Adam, Rychly Pavel, Smrz Pavel and Tugwell David, 2004, « The Sketch Engine », *Euralex*, 2004.

Kittur Aniket and Kraut Robert E., 2008, « Harnessing the wisdom of crowds in wikipedia: quality through coordination », *ACM Conference on Computer Supported Cooperative Work, CSCW*, 2008, San Diego (Ca), p. 37-46.

Kleiber Georges, 1981, *Problèmes de référence: descriptions définies et noms propres*, Paris, Klincksieck.

—, 1987, *Du côté de la référence verbale : les phrases habituelles*, Berne/New York, Peter Lang.

—, 1990, *La sémantique du prototype: catégories et sens lexical*, Paris, Presses universitaires de France.

—, 1994, *Nominales: Essais de sémantique référentielle*, Paris, Armand Colin.

—, 1995, « Sur la Définition des Noms Propres : une Dizaine d'Années Après », *Nom propre et nomination*, Michelle Noailly ed., Paris, Klincksieck, p. 11-36.

—, 1996, « Noms propres et noms communs : un problème de dénomination », *Meta*, vol. 41, no. 4, p. 567-589.

—, 1999a, *Problèmes de sémantique : la polysémie en questions*, Lille, Presses universitaires du Septentrion.

—, 1999b, « Anaphore associative et relation partie-tout : condition d'aliénation et principe de congruence ontologique », *Langue française*, vol. 122, no. 1, p. 70-100.

Koffka Kurt, 1935, *Principles of Gestalt psychology*, London, Routledge & Kegan Paul.

Kosseim Leila and Poibeau Thierry, 2001, « Proper Name Extraction from Non-Journalistic Texts. », *Computational Linguistics in the Netherlands, CLIN'01*, 2001, p. 144-157.

Krenn Brigitte and Evert Stefan, 2001, « Can we do better than frequency? A case study on extracting PP-verb collocations », *ACL Workshop on Collocations*, 2001, Toulouse, France, p. 39-46.

Kuroda Kow, Utiyama Masao and Isahara Hitoshi, 2006, « Getting deeper semantics than Berkeley FrameNet using MSFA », LREC, 2006, Gènes, p. 2425-2430.

Lafferty John D., McCallum Andrew and Pereira Fernando C. N., 2001, « Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data », *18th International Conference on Machine Learning*, 2001, San Francisco (Ca) (ICML '01), p. 282–289.

Lakoff George, 1987, *Women, fire, and dangerous things. What categories reveal about the mind*, Chicago, University of Chicago Press.

Laks Bernard, 2008, « Pour Une Phonologie De Corpus », *Journal of French Language Studies*, vol. 18, no. 01, p. 3-32.

Lamel Lori, Rosset Sophie, Gauvain Jean-Luc, Bennacef Samir, Garnier-Rizet Martine and Prouts Bernard, 2000, « The LIMSI ARISE system », *Speech Communication*, vol. 31, no. 4, p. 339–353.

Langacker Ronald, 1987, *Foundations of cognitive grammar*, Stanford (Ca), Stanford University Press.

Larsson Björn, 2008, « Le sens commun ou la sémantique comme science de l'intersubjectivité humaine », *Langages*, vol. 170, , p. 28.

Laurent Dominique, Nègre Sophie and Séguéla Patrick, 2009, « L'analyseur syntaxique Cordial dans Passage », TALN, 2009, Senlis.

Lebart Ludovic and Salem André, 1994, *Statistique textuelle*, Paris, Dunod.

Lecolle Michelle, Paveau Marie-Anne and Reboul-Touré Sandrine, 2009, *Le nom propre en discours*, Paris, Presses Sorbonne Nouvelle.

Leech Geoffrey, 1992, « Corpora and theories of linguistic performance », *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, J. Svartvik ed., Berlin/New York, Mouton de Gruyter, p. 105–122.

Legallois Dominique, 2004, « Cohésion lexicale et réseaux phrastiques dans la constitution du texte expositif », *L'Unité texte*, S. Porhiel and D. Klingler eds., Paris, Association Perspectives, p. 171-201.

Lehnert Wendy, Cardie Claire, Fisher David, Riloff Ellen and Williams Robert, 1991, « University of Massachusetts: description of the CIRCUS system as used for MUC-3 », *3rd conference on Message understanding*, 1991, Stroudsburg (Pa) (MUC3 '91), p. 223–233.

Lin Dekang, 1998a, « Automatic Retrieval and Clustering of Similar Words », *Coling*, 1998, Montréal (Ca), p. 768-774.

—, 1998b, « Automatic retrieval and clustering of similar words », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, 1998, Stroudsburg, PA, USA, Association for Computational Linguistics (ACL '98), p. 768–774.

Liu Jun and Ram Sudha, 2011, « Who does what: Collaboration patterns in the wikipedia and their impact on article quality », *ACM Transactions on Management Information Systems*, vol. 2, no. 2, p. 11:1–11:23.

Louw Bill, 1993, « Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies », *Text and technology : in honour of John Sinclair*, M. Baker, G. Francis and E. Tognini-Bonelli eds., Philadelphia, Benjamins, p. 229-241.

Léon Jacqueline, 2007, « Meaning by Collocation: The Firthian Filiation of Corpus Linguistics », *History of Linguistics 2005, Selected papers from the Tenth International Conference on the History of the Language Sciences (ICHOLS X)*, vol. 112, , p. 404-415.

—, 2008, « Aux sources de la « Corpus Linguistics » : Firth et la London School », *Langages*, vol. 171, , p. 12-33.

Magnini Bernardo, Negri Matteo, Prevete Roberto and Tanev Hristo, 2002, « A WordNet-based approach to Named Entities recognition », *workshop on Building and using semantic networks*, 2002, Stroudsburg (Pa) (SEMANET '02), p. 1–7.

Mann William C. and Thompson Sandra A., 1988, « Rhetorical Structure Theory: Toward a functional theory of text organization », *Text*, vol. 8, no. 3, p. 243-281.

Manning Christopher D., 2003, « Probabilistic Syntax », *Probabilistic Linguistics*, R. Bod, J. Hay and S. Jannedy eds., Cambridge (Ma), MIT Press, p. 289-341.

Marcu Daniel, 2000, « The rhetorical parsing of unrestricted texts: a surface-based approach », *Computational Linguistics*, vol. 26, no. 3, p. 395–448.

Markert Katja and Nissim Malvina, 2003, « Corpus-Based Metonymy Analysis », *Metaphor and Symbol*, vol. 18, , p. 175-188.

—, 2007, « SemEval-2007 task 08: metonymy resolution at SemEval-2007 », *4th International Workshop on Semantic Evaluations*, 2007, Stroudsburg (Pa) (SemEval '07), p. 36–41.

Martin James R., 2000, « Beyond exchange: APPRAISAL systems in English », *Evaluation in text: authorial stance and the construction of discourse*, S. Hunston and G. Thompson eds., Oxford, Oxford University Press, p. 142-175.

McDonald David D., 1996, « Internal and external evidence in the identification and semantic categorization of proper names », *Corpus Processing for Lexical Acquisition*, B. Boguraev and J. Pustejovsky eds., Cambridge (Ma), MIT Press, p. 21-39.

Mel'cuk Igor A., 1998, « The meaning-text approach to the study of natural language and linguistic functional models », *Forum of the Linguistic Association of Canada and the United States (LACUS)*, 1998, Toronto (Ca), p. 5-19.

Mertens Piet, 2010, « Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE. Conversion vers un format utilisable en TAL », TALN, 2010, Montréal (Ca).

Messiant Cédric, 2009, *Acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts*, Thèse de doctorat, Université Paris-Nord, Paris.

Mikheev Andrei, Moens Marc and Grover Claire, 1999, « Named Entity recognition without gazetteers », *EACL*, 1999, Stroudsburg (Pa) (EACL '99), p. 1–8.

Mill John Stuart, 1884, *A system of logic ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation*, London, Longmans.

Moldovan Dan I., Pasca Marius, Harabagiu Sanda M. and Surdeanu Mihai, 2003, « Performance issues and error analysis in an open-domain question answering system », *ACM Transactions on Information Systems*, vol. 21, no. 2, p. 133-154.

Molla Diego, Zaanen Menno van and Smith Daniel, 2006, « Named Entity Recognition for Question Answering », *Australasian Language Technology Workshop*, 2006, Sydney (Au), p. 51-58.

Morin Emmanuel, 1998, « Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes », TALN, 1998, Paris, p. 172-181.

Mota Cristina and Grishman Ralph, 2008, « Is this NE tagger getting old? », *LREC*, 2008, Marrakech (Ma), p. 1196-1202.

Muller Charles, 1969, « La statistique lexicale », *Langue française*, vol. 2, , p. 30-43.

Nadeau David and Sekine Satoshi, 2007, « A survey of named entity recognition and classification », *Lingvisticae Investigationes*, vol. 30, no. 1, p. 3–26.

Nakov Preslav and Hearst Marti, 2006, « Using verbs to characterize noun-noun relations », *In Proc. of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA), Bularia*, 2006, p. 233–244.

Narayanan Srini and Harabagiu Sanda, 2004, « Question answering based on semantic structures », *Proceedings of the 20th international conference on Computational Linguistics*, 2004,

Stroudsburg, PA, USA, Association for Computational Linguistics (COLING '04).

Nastase Vivi and Szpakowicz Stan, 2003, « Exploring noun-modifier semantic relations », *Proceedings of the 5th International Workshop on Computational Semantics*, 2003.

Nazarenko Adeline, 2004, *Donner accès au contenu des documents textuelsAcquisition de connaissances et analyse de corpus spécialisés*, Mémoire d'Habilitation à Diriger les Recherches, Université Paris-Nord, Paris.

Noailly Michèle, 1995, *Nom propre et nomination: actes du colloque de Brest, 21-24 avril 1994*, Paris, Klincksieck.

Nouvel Damien, Antoine Jean-Yves, Friburger Nathalie and Maurel Denis, 2010, « An Analysis of the Performances of the CasEN Named Entities Recognition System in the Ester2 Evaluation Campaign », *LREC*, 2010, Malte.

Nunberg Geoffrey, 1978, *The pragmatics of reference*, Bloomington, Indiana University Linguistics Club.

Ofoghi Bahadorreza, 2009, *Enhancing Factoid Question Answering using Frame semantic-based approaches*, Phd Thesis, Université de Ballarat, Ballarat (Au).

Ogden Charles K. and Richards Ivor A., 1923, *The meaning of meaning : a study of the influence of language upon thought and of the science of symbolism*, London, Kegan Paul, Trench, Trubner.

Pado Sebastian and Pitel Guillaume, 2007, « Annotation précise du français en sémantique de rôles par projection cross-linguistique », *TALN*, 2007, Toulouse, France.

Palmer Harold E., 1933, *Second Interim Report on English Collocations*, Tokyo, Institute for Research in English Teaching.

Partington Alan, 1998, *Patterns and meanings: using corpora for English language research and teaching*, Amsterdam, Benjamins.

Petasis Georgios, Vichot Frantz, Wolinski Francis, Paliouras Georgios, Karkaletsis Vangelis and Spyropoulos Constantine D., 2001, « Using Machine Learning to Maintain Rule-based Named - Entity Recognition and Classification Systems », *39th Annual Meeting on Association for Computational Linguistics*, 9 July 2001, Toulouse, France, Association for Computational Linguistics (ACL '01), p. 426–433.

Pierini Patrizia, 2008, « Opening a Pandora's Box: Proper Names in English Phraseology », *Linguistik Online*, vol. 36, no. 4, p. 43-52.

Poibeau Thierry, 2003, *Extraction automatique d'information : Du texte brut au web sémantique*, Paris, Hermès.

—, 2006, « Dealing with Metonymic Readings of Named Entities », 28th Annual Conference of the Cognitive Science Society (CogSci 2006), 2006, Vancouver (Ca), p. 1962-1968.

—, 2008, *Des mots au texte. Analyse sémantique pour l'accès à l'information*, Habilitation à diriger les recherches, Université Paris-Nord, Paris.

Purnelle Gérald, 1998, « Théorie et Typographie : une synthèse des règles typographiques de la ponctuation », *À qui appartient la ponctuation: actes du colloque international et interdisciplinaire de Liège, 13-15 mars 1997*, J.-M. Defays, L. Rosier and F. Tilkin eds., Louvain-la-Neuve, Duculot, p. 211-221.

Pustejovsky James, 1998, *The generative lexicon*, Cambridge (Ma), MIT Press.

Pustejovsky James, Hanks Patrick W. and Rumshisky Anna, 2004, « Automated Induction of Sense in Context », *Coling*, 23 August 2004, Genève, p. 924–930.

Pustejovsky James, Ježek Elisabetta and Lenci Alessandro, 2008, « Semantic Coercion in Language: Beyond Distributional Analysis », *Italian Journal of Linguistics*, vol. 20, no. 1 (Special issue: From context to meaning: Distributional models of lexicon in linguistics and cognitive science), p. 181-214.

Radev Dragomir, Fan Weiguo, Qi Hong, Wu Harris and Grewal Amardeep, 2002, « Probabilistic question answering on the web », *11th international conference on World Wide Web*, 2002, New York (WWW '02), p. 408–419.

Rastier François, 1989, *Sens et textualité*, Paris, Hachette.

—, 1995, « La sémantique des thèmes ou le voyage sentimental », *L'analyse thématique des données textuelles : l'exemple des sentiments*, F. Rastier ed., Paris, Didier, p. 223-249.

—, 1996, « La Sémantique des textes : Concepts et Applications », *Hermès*, no. 16, p. 15-37.

—, 2001, *Arts et sciences du texte*, Paris, Presses universitaires de France.

—, 2003, « Les valeurs et l'évolution des classes lexicales », *La polysémie ou l'empire des sens : lexique, discours, représentations*, S. Rémi-Giraud and L. Panier eds., Lyon, Presses universitaires de Lyon, p. 39-56.

—, 2004, « Ontologie(s) », *Revue des sciences et technologies de l'information*, vol. 18, no. 1 (Revue d'Intelligence artificielle), p. 15-40.

Rebeyrolle Josette, 2000, *Forme et fonction de la définition en discours*, Thèse de doctorat, Toulouse II-Le Mirail, Toulouse.

Renouf Antoinette and Sinclair John McH., 1991, « Collocational frameworks in English », *English Corpus Linguistics - Studies in Honour of Jan Svartvik*, K. Aijmer and B. Altenberg eds., London/New York, Longman, p. 128-143.

Riloff Ellen, 1993, « Automatically constructing a dictionary for information extraction tasks », *11th national conference on Artificial intelligence*, 1993 (AAAI'93), p. 811–816.

Riloff Ellen and Jones Rosie, 1999, « Learning dictionaries for information extraction by multi-level bootstrapping », *16th national conference on Artificial intelligence and the eleventh*

Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, 1999, Menlo Park (Ca) (AAAI '99/IAAI '99), p. 474–479.

Rosch Eleanor, Mervis Carolyn B., Gray Wayne D., Johnson David M. and Boyes-Braem Penny, 1976, « Basic objects in natural categories », *Cognitive Psychology*, no. 8, p. 382-439.

Rosset Sophie, Galibert Olivier, Illouz Gabriel and Max Aurélien, 2005, « Interaction et recherche d'information : le projet RITEL », *TAL*, vol. 46, no. 3, p. 155-179.

Sabah Gérard, 2010, « Natural Language Understanding, Where Are We Going? Where Could We Go? », *The Computer Journal*, vol. 54, , p. 1505-1513.

Sager Naomi, 1982, « Syntactic Formatting of Science Information », *Sublanguage: studies of language in restricted semantic domains*, R. Kittredge and J. Lehrberger eds., Berlin/New York, Walter de Gruyter, p. 9-26.

Saint-Aime Sébastien, Le Pévédic Brigitte, Duhaut Dominique and Shibata Takanori, 2007, « EmotiRob: Companion robot Project », *IEEE International Symposium on Robots and Human Communications RO-MAN*, 2007, p. 919-924.

Santini Marina, 2007, « Characterizing Genres of Web Pages: Genre Hybridism and Individualization », *Hawaii International Conference on System Sciences*, 2007.

de Saussure Ferdinand, 1916, *Cours de Linguistique Générale*, Paris, Payot.

Scott Mike and Tribble Christopher, 2006, *Textual patterns: key words and corpus analysis in language education*, Amsterdam, Benjamins.

Sekine Satoshi and Nobata Chikashi, 2004, « Definition, dictionaries and tagger for Extended Named Entity Hierarchy », *LREC*, 2004, Libonne, p. 1977-1980.

Sekine Satoshi, Grishman Ralph and Shinnou Hiroyuki, 1998, « A Decision Tree Method for Finding and Classifying Names in Japanese Texts », *6th Workshop on Very Large Corpora*, 1998, Montréal (Ca).

Sekine Satoshi, Sudo Kiyoshi and Nobata Chikashi, 2002, « Extended Named Entity Hierarchy », *LREC*, 2002, Iles Canaries, p. 1818–1824.

Shannon Claude E., 1948, « The mathematical theory of communication », *Bell System Technical Journal*, vol. 27, , p. 379-423, 623-656.

Silberztein Max, 1993, *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris, Masson.

Sinclair John McH., 1966, « Beginning the study of lexis », *In Memory of J.R. Firth*, C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins eds., London, Longman, p. 410-431.

—, 1987, « Grammar in the dictionnary », *Looking up : an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary*, J. M. Sinclair ed., London, Collins ELT, p. 104-115.

—, 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

—, 2004, *Trust The Text : Language, Corpus and Discourse.*, London, Routledge.

Sinclair John McH., Jones Susan, Daley Robert and Krishnamurthy Ramesh, 2004, *English collocation studies: the OSTI report*, London, Continuum.

Siniakov Peter, 2008, *GROPUS - an adaptive rule-based algorithm for information extraction*, Thèse de doctorat, Université Libre de Berlin, Berlin.

Smadja Frank, 1993, « Retrieving collocations from text: Xtract », *Computational Linguistics*, vol. 19, no. 1, p. 143–177.

Srihari Rohini and Li Wei, 2000, « A question answering system supported by information extraction », *6th conference on Applied natural language processing*, 2000, Stroudsburg (Pa) (ANLC '00), p. 166–172.

Stubbs Michael, 1996, *Text and Corpus Analysis: Computer Assisted Studies of Language and Institutions*, Oxford, Blackwell.

Sundheim Beth M., 1991, « Overview of the third message understanding evaluation and conference », *MUC*, 1991, p. 3-16.

—, 1995, « Overview of results of the MUC-6 evaluation », *MUC-6*, 1995, p. 13-31.

Le Tallec Marc, Villaneau Jeanne, Antoine Jean-Yves, Savary Agata and Syssau Arielle, 2010, « Emologus: a compositional model of emotion detection based on the propositional content of spoken utterances », *13th international conference on Text, speech and dialogue*, 2010, Berlin/Heidelberg, Springer-Verlag (TSD'10), p. 361–368.

Tesnière Lucien, 1965, *Elément de syntaxe structurale*, Paris, Klincksieck.

Tjong Kim Sang Erik F. and De Meulder Fien, 2003, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », *CoNLL*, 2003, Edmonton (Ca).

Tognini-Bonelli Elena, 2001, *Corpus linguistics at work*, Amsterdam, Benjamins.

Tolone Elsa, 2011, *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*, Thèse de doctorat, Université Paris Est, Paris.

Utaker Arild, 1996, « Le problème philosophique du son chez Ferdinand de Saussure et son enjeu pour la philosophie du langage », *Les Papiers du Collège international de philosophie*, no. 23, p. 41-57.

Van Valin Jr. Robert D., 1999, « Generalized Semantic Roles and the Syntax-Semantics Interface », *Empirical Issues in Formal Syntax and Semantics*, F. Corblin, D. C. Sorin and J. M. Marandin eds., La Haye, Thesus, p. 373–389.

Villaneau Jeanne, 2003, *Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée : approche logique pour la compréhension de la parole*, Thèse de doctorat, Université

Bretagne Sud, Lorient/Vannes.

Villaneau Jeanne, Rosset Sophie and Galibert Olivier, 2007, « Semantic Relations for an Oral and Interactive Question-Answering System », *Semantic Representation of Spoken Language (SRSL7)*, 2007, Universidad de Salamanque (Es), p. 12–19.

Voorhees Ellen M., 2000, « Overview of the TREC-9 Question Answering Track », *TREC*, 2000.

Washtell Justin and Markert Katja, 2009, « A Comparison of Windowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations », *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, August 2009, Singapore, Association for Computational Linguistics, p. 628–637.

Wierzbicka Anna, 1992, *Semantics, Culture, and Cognition: Universal Human Concepts in Culture-Specific Configurations*, Oxford/New York, Oxford University Press.

Wilks Yorick, 1975, « A preferential pattern-seeking semantics for natural language inference », *Artificial Intelligence*, vol. 6, , p. 53-74.

Williams Geoffrey, 2006, « La linguistique et le corpus : une affaire prépositionnelle », *Texto*, XI, no. 3-4 (Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du colloque international d'Albi, juillet 2006), p. 151-158.

Wittgenstein Ludwig, 1953, *Philosophical investigations : the German text, with a revised English translation*, 3rd ed., Oxford/Malden (Ma), Blackwell.

Wray Alison, 2002, *Formulaic language and the lexicon*, Cambridge, Cambridge university press.

Yousfi-Monod Mehdi, 2007, *Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus*, Thèse de doctorat, Université Montpellier II, Montpellier.

Zweigenbaum Pierre, Grau Brigitte, Ligozat Anne-Laure, Robba Isabelle, Rosset Sophie, Tannier Xavier, Vilnat Anne and Bellot Patrice, 2008, « Apports de la linguistique dans les systèmes de recherche d'informations précises », *RFLA*, XIII, no. 1.

Indexs

Index des Auteurs

Index Lexical

ANNEXES

Liste des entités-R utilisés dans ce mémoire

Catégories morpho-syntaxiques		
Catégorie	**exemples**	**Définition**
_subs	le <_subs> tabouret </_subs>	substantif
_NN	une <_NN> brosse à dents </_NN>	composé nominal
_action	il <_action> mange </_action>	verbe (infinitif, temps simple, participe passé)
_aux	il <_aux> a </_aux> mangé	auxilliaire
_det	il a changé <_det> la </_det> roue de la voiture	déterminant
_adjectif	on s' expose à l' agression des bandes-annonces <_adjectif> racoleuses </_adjectif>	adjectif
_stat_objet_plus	le <_stat_objet_plus> plus grand </_stat_objet_plus> bateau	adjectif superlatif
_range_objet	le <_range_objet> dernier </_range_objet>	un rang (ordinal, premier, dernier)
_adv	il lit <_adv> calmement </_adv>	adverbe
_adv_compos	<_adv_compos> par rapport à </_adv_compos> hier ...	locution adverbiale
_conjc	il fait beau <_conjc> et </_conjc> chaud	conjonction de coordination
_conjs	il fait beau <_conjs> puisqu' </_conjs> il y a du soleil	conjonction du subordination
_gerondif	il est tombé en <_gerondif> marchant </_gerondif>	participe présent
_prep	il est parti <_prep> depuis </_prep> 3 heures	préposition
_punct	que veux-tu <_punct> ? </_punct>	signe de ponctuation
_modal	ils <_modal> peuvent </_modal> suggérer	verbes modaux
_pronom	pour <_pronom> quiconque </_pronom> ; <_pronom> il </_pronom> a gagné...	les pronoms (personnels, définis, indéfinis, démonstratifs, relatifs etc.)
_mot_inconnu	NA	un mot que le système n'a pas pu taguer

Tags de questions		
Catégorie	**exemples**	**Définition**
_Qqui	<_Qqui> qui </_Qqui> a écrit ...	pronom interrogatif qui
_Qou	<_Qou> où </_Qou> se trouve Paris	formes interrogatives sur lieu
_Qdebdate	<_Qdebdate> depuis quand </_Qdebdate> <_det> la </_det> ...	date de début
_Qfindate	<_Qfindate> quand finira </_Qfindate> <_det> la </_det> ...	date de fin
_Qquand	<_Qquand> quand </_Qquand> Thomas Mann ...	date
_Qannee	<_Qannee> en quelle année </_Qannee> ...	année
_Qheure	à <_Qheure> à quelle heure </_Qheure> a lieu la réunion	heure
_Qqdef	<_Qqdef> qu' est -ce que </_Qqdef> l' ONU	demande de définition
_Qnombre	<_Qnombre> combien </_Qnombre> de pommes ...	question quantificatrice
_Qmaniere	<_Qmaniere> comment </_Qmaniere> est mort M. …	question de manière (*)
_Qpourquoi	<_Qpourquoi> pourquoi </_Qpourquoi> la terre ...	demande d'explications
_Qquel	<_Qquel> quelle est la </_Qquel> forme ...	question générique en quel
_Qquoi	tu manges <_Qquoi> quoi </_Qquoi>	question quoi générique
Tags notionnels/thématiques		
Catégorie	**exemples**	**Définition**
_Tetapesport	la <_Tetapesport> finale </Tetapesport> de rugby	étapes de match en rapport avec le sport
_Tlitterature	un <_Tlitterature> roman </Tlitterature>	concerne la littérature
_Tcinema	un <_Tcinema> film </_Tcinema>	concerne le cinéma
_Ttheatre	<_Ttheatre> une pièce de théâtre </_Ttheatre>	concerne le théatre
_Tcapitale	la <_Tcapitale> capitale </_Tcapitale>	concerne les capitales et villes d'importance
_Tlangue	la <_Tlangue> langue <_Tlangue>	concerne les langues
_Tpopulation	les <_Tpopulation> habitants </_Tpopulation>	concerne la population
_Treligion	la <_Treligion> religion </_Treligion>	concerne les religions
_Tnaiss	la <_Tnaiss> naissance </_Tnaiss>	concerne la naissance
_Tmort	la <_Tmort> mort </_Tmort>	concerne la mort
_Ttransport	le <_Ttransport> bateau	concerne les moyens de transports

	</_Ttransport>	
_Tvocsens	le <_Tvocsens> sens </_Tvocsens>	concerne la définition d'un mot, de son sens
_Tvocety	l' <_Tvocety> origine du mot </_Tvocety>	concerne l'étymologie d'un mot
_Torhographe	le <_Torhographe> masculin </_Torhographe>	concerne l'orthographe d'un mot
Tags dialogiques		
Catégorie	**exemples**	**Définition**
_neg_info	<_Qneg_dial> je ne m' intéresse pas </_Qneg_dial> <_prep> à </_prep> la <_neg_info> <_topic> <_topic_geographie> <_subs> géographie </_subs> </_topic_geographie> </_topic> </_neg_info>	information niée ou non souhaitée
_DDfermeture	<_DDfermeture> au-revoir </_Ddfermeture>	marqueur de fermeture de dialogue
DDouverture	<_DDouverture> bonjour </_Ddouverture>	marqueur d'ouverture de dialogue
_Qdial	<_Qdial> je cherche <1> une </1> information sur le </_Qdial> <_loc> <_pays> Mozambique </_pays> </_loc>	annonce de requête d'information
_Qneg_dial	<_Qneg_dial> je ne veux pas savoir </_Qneg_dial> <_neg_info> <_Qqui> qui </_Qqui> </_neg_info> <_aux> est </_aux> le <_topic> <_topic_geopolitique> <_pers_fonct> <_fonctions> <_fonction> président </_fonction> </_fonctions> de la <_org> <_pays> France </_pays> </_org> </_pers_fonct> </_topic_geopolitique> </_topic>	requête d'information non souhaitée
_DDnon	<_DDnon> pas le </_DDnon>	refus
Tags de personne		
Catégorie	**exemples**	**Définition**
_prenom	<_prenom> Werner </_prenom> K. Rey	les prénoms
_insrt	Werner <_insrt> K. >/_insrt> Rey	les initiales insérées
_partic	Jamil <_partic> al </_partic> Tarifi	les particules proches des noms
_nom	Werner K. <_nom> Rey </_nom>	les noms de famille
_pers_comp	<_pers_comp> <_prenom> Werner </_prenom> <_insrt> K. </_insrt>	un nom de personne complet

	<_nom> Rey </_nom> </_pers_comp>	
_pers	<_pers> <_pers_comp> <_prenom> Jacques </_prenom> <_nom> Chirac </_nom> </_pers_comp>	un nom de personne complet
_annonce_pers	<_annonce_pers> Docteur </_annonce_pers> Edelman	des déclencheurs de personne
_type_pers	<_type_pers> héroïne </_type_pers>	des désignateurs de personne
_Pers	l' <_Pers> <_type_pers> héroïne </_type_pers> </_Pers>	regroupe différents types de désignateurs de personne
_annonce_titre	<_annonce_titre> Mme >/_annonce_titre> Joan Spero	titre de personne
Tags de fonctions		
Catégorie	**exemples**	**Définition**
_fonction_publique	le <_fonction_publique> <1-> premier </1-> ministre </_fonction_publique>	fonctions gouvernementales
_fonction_adm	, <_fonction_adm> directrice générale </_fonction_adm> ,	fonctions dans les entreprises
_fonction_relig	ce <_fonction_relig> pasteur </_fonction_relig> iranien	fonctions religieuses
_fonction_mil	la direction du <_fontion_relig> général </_fonction_relig> Khin	fonctions militaires
_fonction	le <_fonction> président </_fonction> de la France	haute-fonction liée à un lieu
_fonctions	le <_fonctions> <_fonction> président </_fonction> </_fonctions> bosniaque	tag général des fonctions
_pers_fonct	le <_pers_fonct> <_fonctions> <_fonction_adm> chef </_fonction_adm> </_fonctions> <_det> des </_det> <_Organisation> milices </_Organisation> </_pers_fonct>	fonctions dans une organisation
_fp	<_fp> <_pers> <_pers_comp> <_prenom> Jacques </_prenom> <_nom> Chirac </_nom> </_pers_comp> </_pers> <_det> le </_det> <_pers_fonct> <_fonctions> <_fonction> président </_fonction> </_fonctions> <_prep> de </_prep> <_det> la </_det> <_loc> <_pays> France </_pays> </_loc> </_pers_fonct> </_fp>	groupement d'un personne et d'une fonction

_ap	<_fp> <_pers> <_prenom> Antoine </_prenom> </_pers> <_det> le </_det> <_pers_act> <_Apeinture> peintre </_Apeinture> </_pers_act> </_fp>	groupement d'un personne et d'une activité
_fap	le <_fap> <_Ffamille> frère </_Ffamille> <_prep> de </_prep> <_pers> <_prenom> Sophie </_prenom> </_pers> </_fap>	groupement d'un nom de famille et d'une personne
Tags d'organisation		
Catégorie	**exemples**	**Définition**
_orgof	l' <_orgof > <~UN> organisation des nations unies </~UN> </_orgof>	les organisations officielles
_org_medias	le journal économique français la <_org_medias> Tribune </_org_medias>	les medias
_org_peup	le <_org_peup> peuple Kurde </_org_peup>	les peuples
_org_div	<_org_div> Greepeace </_org_div>	organisations diverses
_org_pol	le <_org_pol> parti socialiste </_org_pol>	partis politiques
_syndicat	la <_syndicat> <~CGT> confédération générale des travailleurs </~CGT> </_syndicat>	syndicats
_org_sport	l' <_org_sport> OM </_org_sport>	équipes sportives
_org_univ	l' <_org_univ> Université de <_ville> Genève </_ville> </_org_univ>	les organisation éducatives officielles
_org_relig	les <_org_relig> mulsumans </_org_relig>	Les communautés religieuses
_org_mus	les <_org_mus> Beatles </_org_mus>	les groupes de musique
_org_cinema	<_org_cinama> Miramax >/_org_cinema>	les organisations du cinéma
_org_prob	<_org_prob> <_Org> <_Organisation> office </_Organisation> <_Organisation_genre> fédéral </_Organisation_genre> </_Org> <_prep> de </_prep> <_det> l' </_det> <_subs> environnement </_subs> </_org_prob>	des organisations imprécises

Tags de type d'organisation		
Catégorie	**exemples**	**Définition**
_Organisation	le <_Organisation> société </_Organisation> Truc	déclencheur d'organisation
_Organisation_genre	la <_Organisation> milice </_Organisation> <_Organisation_genre> locale </_Organisation_genre>	modifieur de l'organisation
_Org	<_Org> <_Organisation> chambre </_Organisation> <_det> des </_det> <_typ_org> communes </_typ_org> <_orig> canadienne </_orig> </_Org>	regroupe des tags d'organisation
Tags de type de lieu		
Catégorie	**exemples**	**Définition**
_CCours_d_eau	quel <_CCours_d_eau> fleuve </_CCours_d_eau> traverse Dublin	les types d'étendues d'eau
_MMontagne	quel est la plus haute <_MMontagne> montagne </_MMontagne> de France	Les types de relief géograhique
_PPays	quel est le <_PPays> pays en développement </_PPays> avec ...	Les types de pays
_VVille	quelle est la <_VVille> cité médiéval </_VVille> à côté de Paris	les types de ville
_Astres	une <_Astres> planète </_Astres>	les types d'astres
_LLoc	quel <_LLoc> <_PPays> pays </_Ppays> </_LLoc>	regoupe les tags précédents
Tags de lieux		
Catégorie	**exemples**	**Définition**
_fleuve	la navigation sur le <_fleuve> Rhin </_fleuve>	fleuves, rivères et océans
_adresse	dans le <_adresse> <5-> 5e </5-> arrondissement </_adresse>	arrondissement, numero + rue etc.
_ville	à <_ville> Paris </_ville>	les villes, villages etc.
_province	en <_province> Amérique du sud </_province>	régions, département, états, provinces...
_code_province	<_loc> <_province> Gironde </_province> <_punct> (</_punct> <_code_province> <33> 33 </33> </_code_province> <_punct>) </_punct> </_loc>	code de départements
_pays	la <_pays> France </_pays>	les pays
_montagne	dans les <_montagne> Alpes </_montagne>	les montagnes et sommets
_astres	le <_astres> soleil </_astres>	les planètes, astres etc.

_lieu_remarquable	près d' <_lieu_remarquable> Hollywood </_lieu_remarquable>	lieux remarquables, touristiques etc.
_voie	la <_voie> nationale 215 </_voie>	les rue, route etc.
_coordonnee	le <_coordonnee> 01 69 85 80 02 </_coordonnee>	les coordonnées téléphonique
_monument	le <_monument> château <_prep> de </_prep> Chillon >/_monument>	les monuments
_loc		englobe tous les lieux
_culte	<_culte> église Saint-Pierre </_culte>	les lieux de cultes
_musee	<_musee> musée <_det> des </_det> Arts </_musee>	les musées
_pt_cardinal	au <_pt_cardinal> sud </_pt_cardinal> de l' Europe	les points cardinaux
Tags d'unités de mesure		
Catégorie	**exemples**	**Définition**
_um_superf	3 <_um_superf> hectares </_um_superf>	les unités de superficie
_um_volume	3 <_um_volume> mètres cubes </_um_volume>	les unités de volume
_um_vit	3 <_um_vit> kilomètres / heure </_um_vit>	unités de vitesse
_unitTps	83 <_unitTps> ans </_unitTps>	le temps, la durée
_unitDist	3 <_unitDist> mètres </_unitDist>	les longueurs, distances
_unit_Frq	10 <_unit_Frq> herz </_unit_Frq>	les fréquences
_um_masse	4933 <_um_masse> tonnes </_um_masse>	les masses
_um_div	330 millions de <_um_div> kilowattheures </_um_div>	unités diverses: degrés, intensité électrique, lumins...
_Monnaie	3 <_Monnaie> euros </_Monnaie>	unités monétaires
_Monnaie_subdiv	3 <_Monnaie_subdiv> centimes </_Monnaie_subduv>	les sous-unités monétaires
_val_subs	3 <_val_subs> personnes </_val_subs>	les substantifs précédés d'une valeur
Tags de valeurs stricts		
Catégorie	**exemples**	**Définition**
_gros_val_aine	<_gros_val_aine> <1> une </1> dizaine de millions </_gros_val_aine>	les *aine en millions et milliards
_val_aine	<_val_aine> <1> une </1> vingtaine </_val_aine> de ...	les *aine en-dessous de millions et milliards
_bval	, <_bval> 47 </_bval> ans	les chiffres devant des unités de mesures
_val	<_val> 3 </_val>	les chiffres non suivis d'unités de mesure

_gros_val	<_gros_val> <_val> 2513700 </_val> </_gros_val>	des chiffres >= à 1 million
_pval	<_pval> <_valimp> 105700 </_valimp> </_pval>	les chiffres qui ne sont pas gros
_proportion	<_proportion> <1> un </1> <_prep> sur </_prep> <5> cinq </5> </_proportion>	des expressions de rapport
Tags d'expressions temporelles		
Catégorie	**exemples**	**Définition**
_heure	vers <_heure> <_val_unitTps> <_bval> 14 </_bval> <_unitTps> heures </_unitTps> </_val_unitTps> </_heure>	les horaires
_jour	le <_jour> 13 </_jour> avril	les jours en nombre
_mois	en <_mois> décembre </mois>	les mois
_semaine	le <_semaine> lundi </semaine>	les jours de la semaine
_annee	en <_annee> 1993 </_annee>	les années
_siecle	<_siecle> <12-> 12ème </12-> <_unitTps> siècle </_unitTps> </_siecle>	les siècles
_date_complete	<_time> <_date_complete> <_jour> 30 </_jour> <_mois> juin </_mois> <_annee> 1992 </_annee> </_date_complete>	regroupe _jour/_semaine _mois _annee
_epoque	<_epoque> fin du <_unitTps> millénaire </_unitTps> </_epoque>	période historique
_periode	<_periode> dans l' après-midi </_periode>	les plages d'une journée
_date_relative	<_date_relative> <_det> ce </_det> <_measure_phys> <_unitTps> mois </_unitTps> </_measure_phys> <_adv> -ci </_adv> </_date_relative>	date relative
_age	âgé de <_age> 50 <_unitTps> ans </_unitTps> </_age>	les âges
_time	<_time> <_mois> décembre </_mois> </_time>	regroupe les tags de date
Tags d'événements		
Catégorie	**exemples**	**Définition**
_decl_eve	<_decl_eve> conférence </_decl_eve> ministérielle ...	déclencheur d'événements
_Eve	<_Eve> chute du mur </_Eve> , <_Eve> <_decl_eve> sommet </_decl_eve> de <_ville> Dakar </_ville> </_Eve>	Événement connu

_Eve_sport	<_Eve_sport> <_Tetapesport> finale </_Tetapesport> de la <_champ> coupe du monde </_champ> de <_sports> football </_sports> </_Eve_sport>	Événement sportif
_eve	<_eve> <_range_objet> <28-> 28e </28-> </_range_objet> <_Eve> congrès </_Eve> <_det> du </_det> <_org> PCF </_org> </_eve>	regroupe tous les événements
Tags de récompenses		
Catégorie	**exemples**	**Définition**
_Prix	le <_Prix> Prix Nobel </_Prix>	les types de récompense
_type_prix	<_Prix> Prix Nobel </_Prix> <_type_prix> de la Paix </_type_prix>	le domaine de récompense
_prix	<_prix> <_Prix> Prix Nobel </_Prix> <_type_prix> de la Paix </_type_prix> </_prix>	regroupe les récompenses
Autres tags		
Catégorie	**exemples**	**Définition**
_Tcit	<_Tcit> qui a dit </_Tcit> les poules ont des dents	annonce d'un segment de parole rapportée ou de citation
_cit	<_Tcit> qui a dit </_Tcit> <_cit> les poules ont des dents </_cit>	la parole ou citation proprement dite
_couleur	la <_couleur> blanche </_couleur>	les couleurs
_ langue	le <_langue> suédois </_langue> ...	les langues
_Ffamille	la <_Ffamille> femme </_Ffamille> de Sarkozy	les termes de membres de familles
transport	100 <_transport> Boeing 737 </_transport>	les moyens de transport nommés
_orig	l' armateur <_orig> grec </_orig>	origine de pays

PUBLICATIONS

EL MAAROUF ISMAÏL, 2009, « Natural Ontologies at Work: investigating fairy tales », *Corpus Linguistics Conference*, 2009, Liverpool.

EL MAAROUF ISMAÏL, LE TALLEC MARC & VILLANEAU JEANNE, 2009, « Ontologies naturelles et coercion : formalisation de connaissances à partir d'observations en corpus » *Journées de Linguistique de Corpus,* 2009, Lorient.

EL MAAROUF ISMAÏL, VILLANEAU JEANNE, SAID FARIDA & DUHAUT DOMINIQUE, 2009, « Comparing Child and Adult Language: Exploring semantic constraints », *ICMI MLMI*, 2009, (WOCCI), Boston.

EL MAAROUF ISMAÏL, ROSSET SOPHIE & VILLANEAU JEANNE, 2011, « Extraction de patrons sémantiques appliquée à la classification d'entités nommées », *TALN*, 2011, Montpellier.

EL MAAROUF ISMAIL, 2011, « Verb-Noun Collocations at the crossroads of discourse surface patterns », *Corpus Linguistics Conference*, 2011, Birmingham.