DIC 9410

PROJET DE RECHERCHE

Sophie Piron

Doctorat en Informatique Cognitive
Session d'automne 2003

Table des matières

I. INTRODUCTION

Ce document présente notre projet de thèse. Le titre de la thèse est « Vers un traitement automatique des langues : le cas des verbes de perception sensorielle en français ». La recherche menée combine deux domaines : la linguistique et l'informatique. Par conséquent, la problématique située au centre de cette recherche est double. De plus, même si le thème est linguistique, son analyse dépassera le domaine de la linguistique pour s'inscrire dans le cadre plus large des sciences cognitives. L'utilisation de l'analyse linguistique et cognitive relève du domaine de l'informatique et plus particulièrement du traitement automatique des langues.

Ce document expose d'abord les objectifs de la recherche, ainsi que l'objet linguistique qui la traverse. La deuxième section est consacrée à la description linguistique. Elle traite de la définition des champs sémantiques, de la constitution et de la gestion de ressources linguistiques informatisées. La troisième section de ce document expose les grandes lignes de l'analyse linguistique qui sera développée dans le travail : les restrictions sélectionnelles et la sémantique cognitive. La quatrième section s'attache au formalisme de représentation des données linguistiques, un formalisme à base de contraintes. La cinquième section aborde le traitement automatique des données linguistiques, l'environnement de programmation pour le formalisme de représentation des connaissances choisi. Ensuite, la sixième section identifie les dimensions cognitive et informatique du projet. Finalement, le calendrier de réalisation du projet est donné.

I. 1. OBJECTIFS GÉNÉRAUX

« Le constat primordial établi par le TAL [Traitement Automatique des Langues] est celui de l'importance de la linguistique. Toutes les approches en TAL, lorsqu'elles ne prenaient pas en compte cette discipline dans sa globalité, ont conduit à des impasses. » (Abeillé & Blache 2000, p. 67) Sans nier l'importance des développements accomplis dans le TAL par des systèmes qui font abstraction de l'analyse linguistique (traitements connexionnistes ou probabilistes), nous pensons qu'un système de TAL doit inclure des connaissances linguistiques pour traiter des phénomènes non triviaux comme les extractions, les phénomènes d'accord éloigné, la polysémie, etc.

L'objectif général du TAL est de créer « des programmes informatiques capables de traiter automatiquement les langues naturelles. » (Bouillon *et al*. 1998, p. 5) Dans la perspective que nous adoptons, c'est-à-dire une perspective qui donne à l'analyse linguistique un rôle essentiel, le traitement automatique nécessite trois catégories d'*outils* (Bouillon *et al.* 1998). Il faut en premier lieu des outils linguistiques qui décrivent les diverses connaissances relatives au phénomène de langue étudié. Dans notre cas, il s'agira de développer une analyse des verbes de perception sensorielle en français. En deuxième lieu, des outils formels seront nécessaires pour permettre d'exprimer les connaissances linguistiques dans un formalisme qui convienne au traitement automatique. Pour nous, il s'agira donc de formaliser l'analyse linguistique des verbes de perception sensorielle en français de manière à ce qu'elle puisse être utilisée par un programme. Enfin, en troisième lieu, il est nécessaire de faire appel à des outils informatiques qui

utilisent la description formelle des connaissances dans une application informatique concrète. Dans notre étude, il s'agira d'utiliser la description formelle atteinte dans une application qui procédera automatiquement à une analyse de phrases contenant un des verbes étudiés.

I. 2. SOUS-OBJECTIFS ET DÉGAGEMENT DE LA CONTRIBUTION

Notre contribution relève des différents domaines analysés, mais aussi de leur union. L'analyse linguistique sera fondée sur une perspective historique de la signification des verbes étudiés, qui permettra de mettre en lumière l'évolution sémantique d'une classe de verbes français. Une perspective historique rend possible, d'une part, une meilleure évaluation des frontières entre les champs sémantiques qui composent la signification verbale et, d'autre part, une meilleure compréhension des relations qui relient les différents sens d'un verbe. Cette analyse diachronique donnera un éclairage original sur la façon de traiter les verbes étudiés. De rares études sont faites dans cette perspective. Ainsi, Rastier (2000) a proposé d'étudier l'évolution d'un élément lexical en mettant davantage l'accent sur la valorisation relative des significations que sur leurs relations logiques (extension / restriction) ou sur les figures de rhétorique (métonymie, métaphore,...). Ce volet a pour objectif de démontrer que l'organisation conceptuelle d'un verbe n'est jamais totalement figée, ni même donnée à l'avance. Au contraire, la sémantique se forme et se déforme.

L'analyse linguistique synchronique devrait ensuite permettre de dégager les schémas cognitifs de la perception sensorielle telle qu'elle s'exprime en français dans le cas de certains verbes. Notre analyse devrait compléter certains aspects des analyses déjà existantes ou s'opposer à d'autres aspects de ces analyses. Sans toutes les citer, on relèvera notamment les études de Levin (1993), Viberg (1984), Van Voorst (1992), Van Develde (1977), Akmajian (1977), Brekke (1988), Caplan (1973), Cooper (1974, 1975), Kirsner (1979), Leek & Jong (1982), Rogers (1971, 1972), Quirk (1970), Felser (1999). L'objectif est de dégager les prototypes de sens à l'intérieur du domaine de la perception sensorielle telle qu'elle est exprimée en français. L'analyse devra aussi expliquer les possibilités de polysémie à l'intérieur de ce champ sémantique. Par ailleurs, il s'agira d'étudier la correspondance entre les constructions syntaxiques et les sens pouvant y être attachés. Dans cette optique, il s'agira d'appliquer l'analyse de Levin (1993) au français pour le cas des verbes de perception sensorielle, tout en en approfondissant cette analyse.

Pour étudier la correspondance entre sens et constructions, nous nous baserons sur un corpus que nous avons construit, de telles données rassemblées n'étant pas disponibles. Les différentes annotations introduites dans ce corpus en font une source d'information très utile pour l'analyse linguistique. La formule empiriste de l'analyse sera complétée par des données de nature intuitive (tradition rationaliste).

L'analyse contribuera également à exposer les relations qui existent entre langage et cognition dans le cas de l'expression de la perception sensorielle. Talmy (1975, 1988, 2000) propose une approche qui relie les facultés cognitives générales au langage. Dans notre étude, il s'agira d'étudier la physiologie de la perception sensorielle et

son influence sur l'expression verbale de cette perception en français. Ceci nous amènera à préciser le schéma d'analyse de Talmy (2000).

Les résultats de l'analyse linguistique seront transposés dans un formalisme de représentation qui considère les unités lexicales comme des entités combinant essentiellement sémantique et syntaxe. Il s'agit de HPSG (*Head Driven Phrase Structure Grammar*), une grammaire d'unification qui est arrivée à un stade relativement stable dans son évolution théorique (Pollard & Sag 1988, Pollard & Sag 1994, Sag & Wasow 1999). Notre contribution consistera à développer le lexique computationnel sur lequel repose ce formalisme pour qu'il puisse tenir compte de l'analyse linguistique à laquelle nous aurons abouti. Ceci se fera dans la lignée des travaux dans ce domaine (Lascalides & Copestake 1999, Sanfilippo 1993), travaux qui portent essentiellement sur l'anglais.

Enfin, pour le volet d'implantation de ce travail, l'environnement de développement LKB (*Linguistic Knowledge Building*, notamment Copestake 2002) sera utilisé pour produire un lexique computationnel qui permettra de traiter un certain nombre de phrases en français faisant usage des verbes de perception sensorielle étudiés.

I. 3. L'OBJET D'ÉTUDE LINGUISTIQUE, FIL ROUGE DU TRAITEMENT AUTOMATIQUE

Le fil rouge qui unit les trois grandes étapes établies pour un traitement automatique des langues – analyse linguistique, représentation des connaissances dans un formalisme, application faisant usage du formalisme de représentation– est le phénomène linguistique choisi pour ce traitement automatique. Notre étude a pour objet linguistique les verbes de perception sensorielle en français. Ce que l'on entend par *perception sensorielle* désigne les cinq modes sensoriels dont l'être humain dispose pour appréhender le monde qui l'entoure. Ces cinq modalités sont la vue, l'ouïe, le toucher, l'odorat et le goût. Dès lors, les verbes qui permettent d'exprimer la perception sensorielle identifient, d'une façon ou d'une autre, l'expression d'une de ces cinq facultés. Il existe ainsi un ensemble assez vaste de verbes qui répondent à ce critère. On peut citer comme exemples de verbes de perception : *discerner, entendre, réentendre, palper, humer, sentir, respirer, déguster, goûter, savourer, entrevoir, revoir, voir,…* Parmi ces verbes, on distinguera ceux qui conceptualisent la signification générique d'un système sensoriel (par exemple, *voir* représente le sens type pour la vue) de ceux qui lui apposent des nuances sémantiques (*revoir, entrevoir* par rapport à *voir*). Cette distinction permet de tracer une limite très nette entre deux séries de verbes de perception : les verbes génériques et les verbes nuancés. Il n'existe ainsi que quatre verbes de perception génériques : *entendre* (faculté auditive), *voir* (faculté visuelle), *sentir* (faculté olfactive et tactile) et *goûter* (faculté gustative). Le verbe *toucher* est en fait un verbe exprimant le contact (Levin 1993) et non pas la perception tactile; il ne fait donc pas partie des verbes génériques de la perception. Donc, seuls quatre verbes font l'objet de la présente étude : *entendre, voir, sentir et goûter*.

Le choix de ces verbes a été guidé par l'intérêt cognitif qu'ils représentent, par leur forte polysémie, et par leur usage très courant. On peut espérer que la méthode d'analyse qui leur sera appliquée puisse être étendue aux verbes de leur classe et qu'elle serve de modèle à l'analyse d'autres classes de verbes.

II. DESCRIPTION LINGUISTIQUE

II. 1. PROBLÉMATIQUE GÉNÉRALE ET OBJECTIFS

Le problème essentiel que rencontre le TAL est l'ambiguïté des langues naturelles. Elle peut se révéler à différents niveaux (Bouillon *et al.* 1998) : lexical (de type catégoriel[1] ou sémantique[2]), syntaxique[3], sémantique[4] ou pragmatique[5]. Dans notre travail, nous nous intéressons aux ambiguïtés lexicales de type sémantique (en d'autres termes à la polysémie) pour les verbes *voir, entendre, sentir* et *goûter*, c'est-à-dire aux différents sens que ces verbes peuvent prendre selon le contexte dans lequel ils apparaissent. Nous verrons qu'un verbe impose des traits de sélection sur les mots qui l'entourent, qui, à leur tour, impliquent un choix dans les sens possibles du verbe. Il s'agit d'un champ sémantique qui se crée autour du verbe (Lyons 1977, Allen 1987, Smith 1991). Par exemple, dans la phrase *j'ai entendu un bruit dans la pièce d'à côté,* le verbe *entendre* impose d'abord à son contexte des traits de sélection : il peut, entre autres, accepter des mots représentant des phénomènes auditifs. Dans l'exemple ci-dessus, la phrase est sémantiquement acceptable grâce à la présence de mots respectant ce trait de sélection. En même temps, le verbe est orienté vers le sens de la perception auditive par la cooccurrence du phénomène auditif *bruit* et de la localisation ce phénomène auditif (*dans la pièce d'à côté*). Nous aborderons aussi l'ambiguïté syntaxique pour les cas de rattachement des syntagmes prépositionnels (*dans la pièce d'à côté*).

La polysémie des verbes de perception sensorielle tient à plusieurs facteurs. Ces verbes sont non seulement très courants, mais ils sont aussi très centraux dans le langage de par les phénomènes perceptifs qu'ils expriment. De fait, notre perception et notre conception du monde sont modelées par nos sens. Et nous constaterons que la perception modèle notre compréhension du monde qui nous entoure. Il faut souligner ici que l'extension sémantique des mots ne se fait pas de manière arbitraire (Atkins & Levin 1992; Pustejovsky 1991, 1994; Ostler & Atkins 1992). Des régularités sont décelables par l'analyse linguistique. C'est l'objectif que nous nous fixons : quelles sont les régularités de la polysémie de la perception sensorielle verbale ?

---

[1] Il y a ambiguïté lexicale catégorielle lorsque plusieurs catégories syntaxiques peuvent être attribuées à un mot. Par exemple, « guide » peut être soit le nom commun soit le verbe *guider* conjugué à la 1ere ou à la 3e personne du singulier de l'indicatif présent ou du subjonctif présent.

[2] On parle d'ambiguïté sémantique quand un mot a plusieurs sens possibles : *entendre* peut signifier *ouïr*, *comprendre,* etc. Il est à noter que Bouillon et al. (1998) intègre dans la catégorie de l'ambiguïté la définition de la polysémie.

[3] Un syntagme prépositionnel peut se rattacher par exemple au syntagme qu'il complète (*j'ai entendu un bruit dans la cuisine*) ou au verbe (*j'ai entendu un bruit dans la cuisine* [*j'étais dans la cuisine*]; *j'ai entendu un bruit de la cuisine* [*depuis la cuisine*]).

[4] Bouillon et al. (1998) mettent dans ce type d'ambiguïtés notamment les ambiguïtés découlant de la portée des quantificateurs : *tous les étudiants ont entendu un bruit*.

[5] Il s'agit des données provenant du contexte, notamment la référence des pronoms : *il a entendu un bruit parce qu'il avait enlevé ses écouteurs, il a entendu un bruit parce qu'il s'est produit dans la pièce d'à côté*.

II. 2. LES CHAMPS SÉMANTIQUES

II. 2.1. Problématique et objectifs

Dans l'aspect plus proprement linguistique de cette étude, la première étape relève du domaine de la diachronie. À ce sujet, Talmy (2000) insiste sur le fait que les analyses reposant sur l'introspection doivent être complétées par des analyses provenant d'autres méthodologies : « the analysis of discourse and corpora, crosslinguistic and diachronic analysis […] and the instrumental probes of neuroscience. » (p. 5). Le but que nous poursuivons s'inscrit grandement dans la proposition de Talmy. Dans cette section, nous exposons l'objectif qui consiste à étudier l'évolution sémantique de chaque verbe de perception sensorielle, c'est-à-dire à décrire les différentes significations que ces verbes ont prises au cours du temps. Lorsque l'on travaille sur la sémantique verbale, on se trouve très vite devant des significations qui semblent éloignées les unes des autres ou, du moins, des sens qui additionnent des traits sémantiques déjà nuancés. Par conséquent, les liens entre ces significations apparaissent parfois flous. Mettre au jour les sens perdus aide à comprendre et à présenter l'aspect actuel des champs sémantiques couverts par un verbe. En effet, de cette manière, il est possible de mettre en lumière des liens sémantiques entre plusieurs significations, liens que l'évolution peut avoir masqués. Par ailleurs, une analyse diachronique fait apparaître plus nettement les champs puisqu'elle permet de compléter et d'affiner le réseau sémantique d'un verbe. Il s'agit également de mettre en avant que l'expression de la perception sensorielle ne va pas de soi.

Une analyse diachronique du lexique se doit d'établir la différence entre les changements onomasiologiques et sémasiologiques. Les premiers sont des changements qui portent sur la dénomination d'un concept. Ainsi, dans le cas des verbes de perception un changement onomasiologique a-t-il eu lieu au XVIIe siècle : la perception auditive ne s'est plus exprimée au moyen du verbe *ouïr*, mais bien du verbe *entendre*. Quant aux changements sémasiologiques, ils relèvent de différences dans les significations. Par exemple, le verbe *entendre* a subi de tels changements lorsqu'il a perdu le sens de *tendre, étendre* vers le XVIe siècle.

II. 2.2. Méthodologie

La consultation d'ouvrages lexicographiques de référence sur les divers états du français a permis de tracer l'évolution sémantique des verbes qui nous occupent. La description de l'évolution ne se base que sur les notices des dictionnaires. Par ailleurs, il faut bien accepter comme telles les informations chronologiques proposées par les dictionnaires.

Le dictionnaire choisi pour représenter les états les plus anciens de la langue est le *Dictionnaire de l'ancienne langue française et de tous ses dialectes, du IXe au XVe siècle* de Godefroy. Cet ouvrage couvre une très longue période

chronologique puisqu'il décrit le français depuis le IXe siècle jusqu'au XVe siècle. Il dépouille ainsi les usages de l'ancien français classique (1150-1350), mais aussi du moyen français (1350-1500)[6].

Le français de la Renaissance est représenté, dans cette étude, par deux dictionnaires. Le premier est le *Dictionnaire de la langue française du XVIe siècle* de Huguet. Le second ouvrage choisi pour représenter cette période est le *Thresor de la langue francoyse tant ancienne que moderne* (1606) de Nicot et Rançonnet. Cet ouvrage est un remaniement de l'édition de Thierry (1564) de l'ouvrage initialement écrit par Robert Estienne (*Le dictionnaire Francois latin*, 1539-1540). L'édition de Nicot et Rançonnet propose des articles considérablement augmentés par rapport à l'édition précédente. C'est un dictionnaire qui présente les mots français et leurs équivalents latins, avec des explications en latin.

Deux éditions du *Dictionnaire de l'Académie française* ont été choisies pour représenter le français classique. La première édition du *Dictionnaire* (1694) représente l'usage de la langue du XVIIe siècle. L'autre édition choisie pour représenter le français classique est la cinquième édition du *Dictionnaire de l'Académie française* (1798). Elle permet de représenter la langue du XVIIIe siècle. Le choix de cet ouvrage fut dicté par la transition qu'il opère au sein des différentes éditions du *Dictionnaire*. En effet, le bouleversement politique de la Révolution a transformé la vision initiale d'une description puriste du bon usage en une description plus réformiste. C'est cette vision qui confère à la cinquième édition du *Dictionnaire*, publiée à la toute fin du XVIIIe siècle, un caractère de représentation assez fidèle de l'état de la langue au cours de ce siècle.

La sémantique que les verbes de perception ont couverte au cours du XIXe siècle a été établie sur la base du dépouillement des notices de quatre dictionnaires. Il s'agit d'abord de l'ouvrage de référence qu'est le Littré (*Dictionnaire de la langue française*); ensuite, du *Dictionnaire complet illustré de la langue française* de P. Larousse et, enfin, du Dictionnaire de l'Académie française (éditions de 1835 et de 1877). Ces quatre dictionnaires présentent l'état de la langue tout au long du XIXe siècle. Enfin, le français actuel est représenté par *Le Robert électronique* (1994), *Le grand usuel Larousse* (1997) et le *Trésor de la langue française* (1971 – 1994). L'ensemble des ouvrages lexicographiques consultés couvre donc les différentes époques du français.

La compilation des notices de tous ces dictionnaires est à la base de l'analyse diachronique proposée (Piron 2002a, b). Les définitions données par les différents ouvrages lexicographiques ont été assemblées pour être ensuite regroupées en champs sémantiques. Chaque verbe étudié dans le cadre de cette recherche présentera donc une section en sémantique diachronique. Il s'agira d'y exposer les significations prises au cours des siècles, depuis l'ancien français jusqu'au français actuel, de manière à proposer la carte sémantique de chacun de ces verbes.

---

[6] Il existe d'autres ouvrages lexicographiques pour cette période. Ils n'ont pas été retenus ici. Dans le *Tobler et Lommatzsch,* les définitions sont sommaires bien que les exemples soient nombreux. Le FEW (***Französisches Etymologisches Wörterbuch***) est moins facile à consulter, car il propose un classement par étymons, sans index général. Quant au Greimas (*Dictionnaire de l'ancien français*), il s'agit d'un dictionnaire d'usage courant. Il est donc normal qu'il présente des lacunes, tant dans le classement des sens, que dans les datations.

II. 3. CARTE CONCEPTUELLE DES VERBES

Dans le domaine de la sémantique lexicale, les études se fondent sur quatre opérations pour qualifier les changements de sens : l'extension et la restriction, la métaphore et la métonymie. Ces opérations se regroupent d'ailleurs en deux catégories : d'une part, l'extension et la restriction relèvent des opérations logico-référentielles (logique de classes, qui décrit des relations d'inclusion) ; d'autre part, la métaphore et la métonymie relèvent de la rhétorique traditionnelle (théorie des tropes[7]). Ces opérations servent à décrire tant des relations synchroniques (Martin 1992) que des relations diachroniques (citons seulement Darmesteter 1887).

Il est possible cependant de poser le problème de l'évolution sémantique dans son ensemble sans avoir recours à la logique ou à la rhétorique. D'ailleurs, le but de notre étude en sémantique diachronique n'est pas de proposer une analyse approfondie des facteurs des changements sémantiques, mais simplement de rendre compte des variations sémantiques. Nous proposons donc une présentation de l'évolution en nous basant sur les champs qui découpent l'espace sémantique d'un verbe.

L'historique des acceptions du verbe *entendre* (Piron 2002a, 2002b, 2003), par exemple, permet de constater que les significations que nous connaissons aujourd'hui n'existent pas toutes depuis l'origine du verbe. Dans cette optique, il faut souligner que les emplois du verbe *entendre* signifiant l'attention mentale (*s'occuper de*, par exemple), la compréhension et la connaissance sont particuliers au français, car dans cette langue les deux verbes latins *intendere* (qui a donné *entendre*) et *audire* (qui a donné *ouïr*) se sont confondus avec le temps. À l'inverse, dans une autre langue romane, en espagnol, les deux verbes latins ont donné deux verbes séparés : *audire* a donné le verbe *oír,* qui signifie « percevoir par l'ouïe » et *intendere* a donné *entender*, qui signifie « comprendre ». Dans le cas de l'anglais, une langue germanique, le verbe de la perception auditive, *to hear*, ne présente que les acceptions relatives à l'audition : *entendre; écouter; écouter (dans le contexte juridique); exaucer; entendre dire* et *recevoir des nouvelles.*

Le graphique suivant permet de visualiser les acceptions du verbe français *entendre* au cours du temps. Les périodes passées en revue sont représentées en abscisse tandis que les différentes significations apparaissent en ordonnée, classées par champs sémantiques. Il y a trois champs : le champ sémantique physique (1), sensoriel (2) et mental (3).

[7] Il est étonnant de constater que la rhétorique fait exclusivement usage de la métaphore et de la métonymie en excluant les autres tropes. De plus, la métonymie n'est qu' « un cas particulier de l'extension » (Rastier 2000, p. 142).

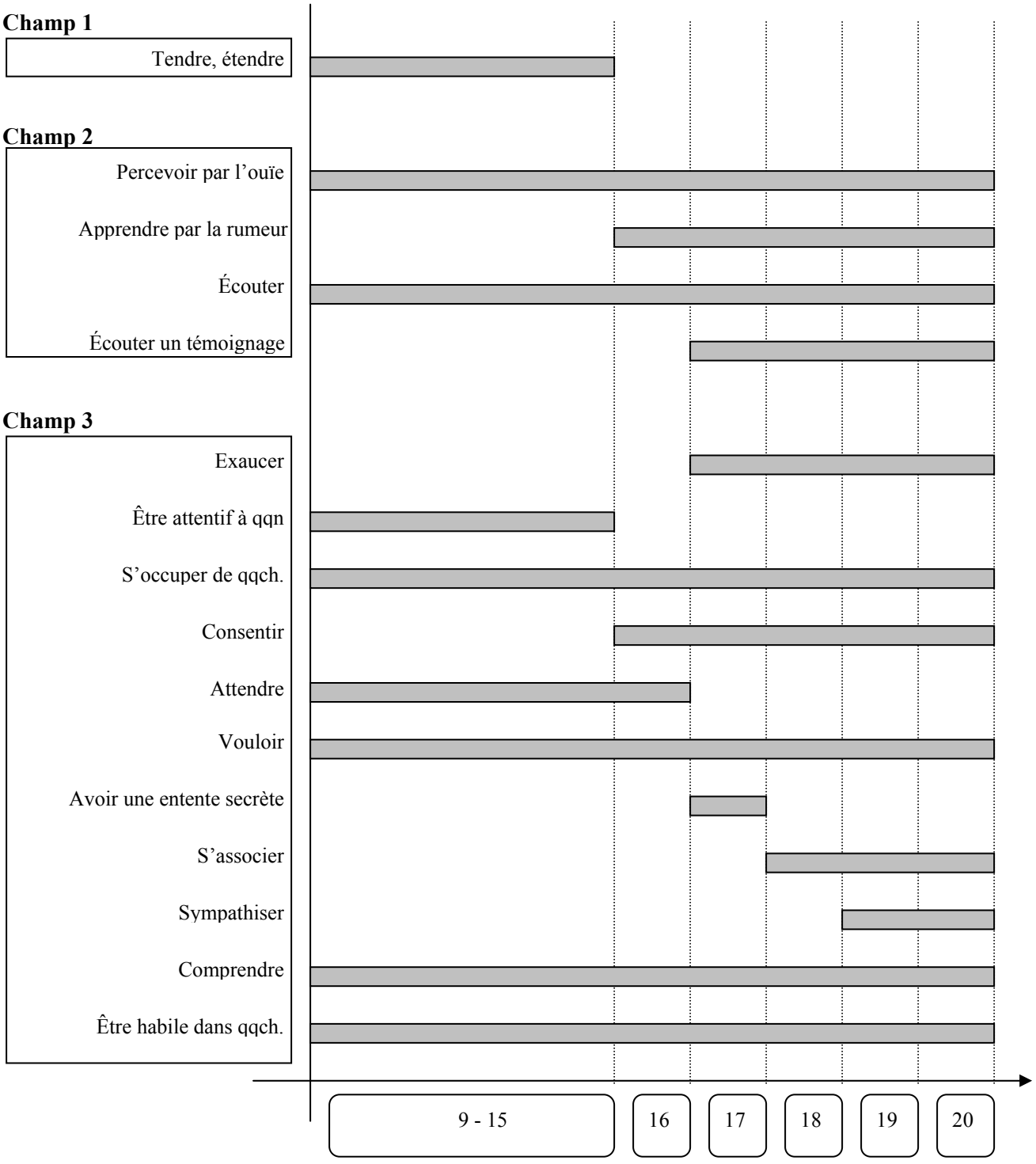

Figure1. Les acceptions du verbe *entendre* au cours des siècles.

Notre travail exposera les évolutions sémantiques de chacun des verbes étudiés pour définir leurs espaces sémantiques respectifs. La brève description du cas de *entendre* que nous avons présentée ici a permis de dégager l'aspect général de sa sémantique et cela sert de base conceptuelle à l'analyse sémantique synchronique.

II. 3. CONSTITUTION ET GESTION DE RESSOURCES LINGUISTIQUES INFORMATISÉES

II. 3.1. Problématique et objectifs

Après les heures de gloire des études de corpus à la fin du XIXe siècle et au cours de la première moitié du XXe siècle, le discrédit est jeté sur ces études empiriques par le courant chomskyen (Chomsky 1956, 1957). La tradition d'études sur corpus persiste cependant, entre autres pour des analyses sociolinguistiques. Au cours des années 80, le TAL vient à s'intéresser aux corpus grâce aux percées technologiques faites par des méthodes empiriques dans le domaine de la reconnaissance de la parole (Church & Mercer 1993). Depuis cette époque, les corpus annotés sont devenus des outils essentiels tant pour les linguistes que pour les ingénieurs (Véronis 2000). Il faut distinguer la linguistique de corpus –qui analyse les corpus d'un point de vue linguistique– des corpus informatisés, qui sont nés de la nécessité d'avoir des corpus disponibles et qui font l'objet de travaux informatiques. C'est ainsi que deux directions ont été prises par rapport aux corpus : la construction de corpus annotés (le corpus SUSANNE, le corpus Brown, etc.) et l'utilisation de corpus (Lebart et Salem 1994, Stubbs 1996, Sinclair 1997). Pour ce qui est de la construction de corpus annotés, les corpus constituent souvent une fin en soi puisqu'il s'agit de développer des outils complexes de traitement automatique comme l'annotation morphologique, l'annotation syntaxique, etc. Il s'agit là d'un domaine encore en expansion. Pour ce qui est de leur utilisation, les corpus sont utiles pour l'analyse linguistique, pour l'extraction de ressources (des lexiques, par exemple), pour la mise au point et le test d'outils, pour l'apprentissage automatique, etc. Dans le cas de l'analyse linguistique, un corpus sert, en amont, à la modélisation d'une analyse. Il sert ainsi à rechercher des occurrences, des structures syntaxiques, etc. Un corpus sert aussi, en aval, à valider des modèles linguistiques. Actuellement, dans le domaine du TALN, les chercheurs visent « des traitements du langage ancrés fortement dans les données attestées plutôt qu'une analyse en profondeur de domaines restreints. » (Valli 2000, p. 77).

Notre recherche s'inscrit dans cette double optique du TALN (construction – utilisation). Le but poursuivi consiste, dans un premier temps, à bâtir un corpus et, dans un second temps, à utiliser ce corpus essentiellement comme terrain d'analyse linguistique. Les données formant le corpus doivent être encodées dans un format qui permette une bonne gestion de ces ressources linguistiques : il faut pouvoir les classer en fonction de caractéristiques qui leur auront été attribuées, mais il faut aussi que le corpus soit consultable en tant que ressource linguistique.

II. 3.2. Constitution du corpus

L'objectif poursuivi est de constituer un corpus qui serve de terrain d'analyse linguistique : il s'agit d'étudier la variété des emplois des verbes de perception sensorielle. Il faut donc un corpus représentatif pour guider cette analyse.

a. Définition du corpus

« Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » (Sinclair 1996, p.4). Sinclair (1996) distingue deux types de corpus : les corpus de textes et les corpus d'échantillons. Les premiers reprennent des textes complets alors que les seconds ne sont composés que de morceaux de textes, voire de phrases. Les échantillons doivent présenter, autant que possible, une taille similaire. Le corpus que nous avons élaboré appartient à la seconde catégorie, celle des corpus d'échantillons.

Un corpus se définit également par le type de langue qu'il représente. Il y a deux possibilités (Sinclair 1996). La première est celle du corpus spécialisé. Un corpus est dit spécialisé s'il contient uniquement des échantillons relevant d'un seul type de situation de communication ou d'un seul domaine (domaine scientifique, domaine technique, etc.). La seconde possibilité est un corpus de référence. Ce type de corpus vise à représenter la langue générale, dans toutes ses variétés pertinentes. Le corpus que nous avons construit est un corpus dit de référence puisqu'il vise à définir la langue générale[8].

Le type de langue que consignent les corpus peut aussi varier selon qu'il s'agit de langue écrite ou de langue parlée (dans ce cas, le corpus transcrit des enregistrements). Notre corpus relève du domaine de la langue écrite. Le type de langue varie également selon les registres qui sont représentés dans le corpus. Ainsi, le corpus que nous avons construit présente deux registres : l'un est littéraire; l'autre est un registre courant, de type journalistique.

Finalement, un corpus se définit par l'étendue temporelle qu'il représente. Il existe des corpus qui intègrent des données de différentes périodes. Par exemple, le *Trésor de la Langue Française* ou le *Grand Robert* comprennent des corpus de citations issues de textes s'étalant du XVI^{e} au XX^{e} siècle. Mais de tels corpus ne constituent pas forcément des corpus adaptés aux études diachroniques. En effet, généralement, il n'y a pas suffisamment de représentation pour chaque période chronologique. Au vu de ces problèmes (représentation très inégale des diverses périodes) et du but que nous poursuivons (analyse de la sous-catégorisation verbale dans l'emploi contemporain des verbes de perception), notre corpus est synchronique : il ne comporte que des données appartenant au français contemporain. En conclusion, nous avons donc un corpus d'échantillons, monolingue, général et synchronique.

b. Données brutes

Le registre littéraire de notre corpus est formé de citations provenant du *Robert Électronique*. Nous avons d'abord recueilli toutes les citations comprenant les verbes que nous étudions. Initialement, ces citations appartiennent à des périodes s'étalant du XVI^{e} au XX^{e} siècle. Nous avons alors classé les écrivains par périodes chronologiques à l'aide

[8] Notre choix étant de couvrir l'usage général des verbes de perception, nous avons opté pour un corpus de référence, mais il aurait été possible d'opter pour un corpus spécialisé (par exemple, un corpus médical) pour représenter des emplois plus spécifiques des verbes de perception.

du *Dictionnaire des oeuvres littéraires* de Bordas. Nous n'avons retenu que celles relevant de la période moderne, à savoir le XX^{e} siècle. Quant au registre journalistique neutre, il a été obtenu au moyen du programme de concordances *GlossaNet*. Cet outil permet d'extraire des concordances dans des journaux en ligne et cela, dans plusieurs langues. C'est ainsi que nous avons obtenu des concordances pour les verbes *voir, entendre, etc.* dans le journal belge francophone *La Meuse*. Pour le registre courant, des domaines variés sont donc représentés : la politique, l'économie, la culture, les sports, les sciences et les nouvelles quotidiennes. La période couverte s'étend du mois d'octobre 2001 au début du mois de janvier 2002.

Les corpus standard qui font office de référence présentent de grandes variations dans leur taille : ils vont de seulement quelques milliers à plusieurs millions de mots. Par exemple, le corpus SUSANNE compte 130 000 mots (http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/public/susanne.html) alors que le British National Corpus en compte 100 106 008 (http://www.natcorp.ox.ac.uk/). La taille d'un corpus ne peut véritablement être appréciée qu'en fonction d'autres paramètres, notamment les annotations que l'on ajoute et la manière dont le corpus est traité. Si les annotations sont relativement fines et qu'elles sont ajoutées manuellement, il est normal que l'envergure du corpus soit moindre. Le but que nous poursuivons en faisant une étude de corpus est d'obtenir un nombre d'occurrences des verbes de perception suffisamment important pour étudier leurs emplois. Par conséquent, il ne nous est pas nécessaire de disposer de plusieurs millions de mots pour établir des régularités de constructions et de sens. Pour l'instant, nous ne disposons que des chiffres pour les données relatives au verbe *entendre.* Le corpus comprend 2080 citations, pour un total d'environ 137 000 mots.

II. 3.3. Annotation du corpus

a. Problématique et objectifs

L'objectif poursuivi est de construire un corpus ayant une valeur informationnelle ajoutée qui permette de regrouper les emplois des verbes étudiés. Pour ce faire, des annotations de différents types ont été apposées aux citations de manière à pouvoir repérer des régularités sur l'association entre les types de constructions utilisés pour un verbe et les sens y correspondant.

Un premier balisage du matériel textuel brut a été effectué pour obtenir une ressource dite primaire (Bonhomme 2000). Une référence a été attribuée à chaque citation pour identifier, entre autres choses, son registre: par exemple, entendre_LaMeuse_021001_04, entendre_GrandRobert_cit20_124.

Si la valeur d'un corpus s'apprécie sur la base de trois critères –la taille, la diversité et l'annotation– c'est l'annotation qui est le plus important de tous (Garside *et al.* 1997). C'est en effet ce qui donne le plus de valeur ajoutée à un corpus. L'utilité de l'étiquetage[9] est surtout de pouvoir extraire des informations linguistiques (collocations, choix de

[9] Nous reviendrons plus loin sur la distinction entre annotation et étiquette. Pour l'instant, nous parlons d'annotation et d'étiquetage de manière synonymique.

constructions selon les registres, etc.) en fonction des besoins de l'analyse. L'étiquetage relève d'une analyse linguistique des données, une forme d'interprétation. Les types d'annotations qui existent sont les suivants : annotation phonétique (transcription et prosodie), grammaticale (morphosyntaxe et syntaxe), sémantique (mots et phénomènes discursifs) et multilingue (phrases et mots). L'annotation de notre corpus est grammaticale (de type morphosyntaxique et syntaxique) et sémantique (portant sur les mots). Nous les développons dans les sections suivantes.

La majorité des schémas existants pour l'annotation, notamment pour l'annotation syntaxique, ont été établis pour l'anglais. On relève le Lancaster-Leeds Treebank; le Lancaster Parsed Corpus; le corpus SUSANNE; le Penn Treebank; le schéma ENGCG; le schéma TOSCA; le British National Corpus, etc. De tels travaux se font de plus en plus dans d'autres langues. Ainsi, des systèmes pour le finnois, le suédois, le danois, le néerlandais, l'allemand, l'espagnol, l'italien et le français sont en cours d'élaboration, mais il s'agit souvent d'une extension des schémas utilisés pour l'anglais. Pour le cas du français, le schéma utilisé pour le corpus IBM Paris est parallèle à celui utilisé pour l'annotation de l'anglais dans le corpus Lancaster/IBM treebank. Pour le français, on relèvera aussi la construction d'un corpus arboré par l'équipe d'Abeillé (Abeillé & Clément 1999, Abeillé *et al.* 2001, Abeillé *et al.* 2003). Notre corpus se veut une contribution au mouvement en faveur de travaux basés sur des corpus en langue française, tout en apportant des particularités dans les annotations.

b. Méthode d'annotation

Au plus on augmente la taille d'un corpus, au plus il faut recourir à l'automatisation de l'annotation. Si l'annotation manuelle peut être lente et difficile, il n'y a, jusqu'à présent, aucun système d'étiquetage automatique qui soit totalement fiable. Il y a toujours une marge d'erreur. Les solutions d'automatisation sont donc des solutions intermédiaires. L'étiquetage manuel demeure envisageable pour les corpus de taille moyenne (en deçà du million de mots) et pour autant qu'il n'y ait pas trop d'étiquettes parmi lesquelles choisir. Puisque notre corpus répond à ces deux critères, il permet une annotation manuelle. Parmi les corpus de référence, le corpus SUSANNE, qui est un des plus petits corpus, a été annoté manuellement. De même, le corpus IBM Paris, qui est un corpus en français, a été annoté manuellement par une équipe de linguistes. À l'opposé, l'annotation syntaxique générale de la majorité des corpus se fait interactivement (projet TOSCA : Oostdijk 1991, Van Halteren & Oostdijk 1993) ou par correction manuelle (le Penn Treebank, par exemple).

c. L'annotation grammaticale

Nous décrirons d'abord les annotations grammaticales que l'on rencontre dans les corpus de référence. Par la suite, nous exposerons les choix effectués pour l'annotation de notre corpus.

c.1. Les corpus de référence

L'annotation POS (*Part Of Speech*) ou annotation morphosyntaxique consiste à identifier les catégories grammaticales de chaque mot et leurs particularités morphologiques. L'exemple suivant, tiré du corpus français IBM Paris (Langé 1994), « ce_DDEMMS guide_NCOMS », indique au moyen de codes que le mot « ce » est un déterminant démonstratif masculin singulier et que le mot « guide » est un nom commun masculin singulier. Le degré de granularité de telles annotations varie d'un corpus à l'autre. Généralement, l'annotation morphosyntaxique est un prérequis pour l'annotation syntaxique. Les systèmes automatiques l'intègrent donc. Déjà à cette étape, il est nécessaire de faire de la levée d'ambiguïté[10].

Il y a moins de consensus pour l'annotation syntaxique que pour l'annotation morphosyntaxique, notamment quant à la définition des segments. Les annotations syntaxiques sont généralement de type PS (*phrase structure*) et découpent la phrase en syntagmes. De manière générale, les schémas d'annotation présentent une certaine indépendance par rapport aux théories linguistiques. EAGLES (1996) relève 8 types possibles pour l'annotation syntaxique. Ces types sont appelés des couches d'annotation syntaxique : le parenthétisage des segments, l'étiquetage des segments, l'identification des relations de dépendance, les étiquettes fonctionnelles[11], la sous-catégorisation des segments syntaxiques, l'information profonde (dite logique), l'information à propos du rang de l'unité syntaxique (principale, subordonnée) et enfin les caractéristiques syntaxiques spécifiques à la langue parlée.

L'ensemble de ces 8 couches présente une forme de hiérarchie assez relâchée. Cela dit, dans la majorité des cas, les couches que l'on peut qualifier *de base*, tel le parenthétisage ou l'étiquetage des segments, sont nécessaires à l'élaboration des couches suivantes. Mais en même temps, il n'y a pas de précédence stricte d'une couche à l'autre. Si l'on regarde les niveaux d'annotation syntaxique que l'on trouve dans les corpus les plus connus, on constate que, dans le cas du corpus IBM Paris (colonne 4, ci-dessous), seuls les deux premiers niveaux sont présents, c'est-à-dire la segmentation et l'étiquetage des segments.

[10] La levée d'ambiguïté peut être atteinte automatiquement (règles contextuelles, grammaires locales, etc.). Mais, dans certains cas, ce ne sera pas suffisant et il faudra avoir recours au linguiste. La même situation se reproduit avec les annotations syntaxiques.

[11] Si la plupart des théories syntaxiques reposent sur l'analyse en constituants (*phrase structure*), le statut des fonctions grammaticales est plus controversé. Dans les théories de la syntaxe de dépendance, Mel^chuk (1988) pense que la structure fonctionnelle est plus fondamentale que la structure des constituants. L'annotation de la structure fonctionnelle est d'ailleurs devenue plus importante récemment. Quelques schémas d'annotation sont basés sur la théorie syntaxique de la dépendance. Ainsi, par exemple, le Prague Dependency Treebank du tchèque (Hajic 1998) et le METU Treebank pour le turc (Oflazer *et al*. 2000).

Couches	(1)	(2)	(3)	(4)	(5)	(6)
(a) Parenthétisage	☑	☑	☑	☑	☑	-
(b) Annotation	☑	☑	☑	☑	☑	-
(c) Relations de dépendance	-	x*	x*	-	-	☑
(d) Fonctions (sujet, COD, …)	-	☑	☑		☑	☑
(e) Sous-catégorisation	-	☑	-	-	☑	x*
(f) Information logique	-	☑	☑	-	-	x*
(g) Rang	☑	☑	☑	-	☑	-
(h) Langue parlée	-	x*	-	-	x*	-

x* : en cours d'élaboration.

- : pas implémenté

Tableau 1. Panorama des niveaux d'annotation dans certains corpus de référence[12] (extrait de EAGLES 1996).

À cause de la nature très variable des annotations syntaxiques et du fait que des schémas déjà existants n'indiquent pas des couches d'annotations de base (c'est le cas du schéma ENGCG, qui n'indique ni le parenthétisage ni l'étiquetage des segments), le groupe EAGLES n'a considéré aucune annotation syntaxique comme obligatoire. Par contre, des recommandations ont été émises pour les catégories utilisées dans l'étiquetage des segments. Ces catégories s'insèrent dans une annotation appliquant les modèles de structure de phrase (*phrase structure models*). Voici ces catégories recommandées : phrase, clause (subordonnée, sous-phrase), syntagme nominal, syntagme verbal syntagme adjectival, syntagme adverbial et syntagme prépositionnel. Les autres types d'annotations sont considérés optionnels.

L'annotation grammaticale (morphosyntaxe et syntaxe) peut se faire sous trois formats. Une première solution pour représenter les annotations est le parenthétisage sous un format vertical, c'est-à-dire en colonnes. La phrase originale est dans la première colonne, les étiquettes morphosyntaxiques, dans la seconde et les annotations syntaxiques, dans la troisième. L'exemple 1 est tiré de l'Associated Press Corpus (EAGLES 1996). Il est important d'insister une fois encore sur la difficulté de procéder à un étiquetage automatique et sur les nombreuses et longues étapes qui précèdent le produit final.

```
The       AT    [N
door      NN1
,         ,
which     DDQ   [Fr[N]
was       VBDZ  [V
equipped  VVN
with      IW    [P
neither   LE    [N
bell      NN1   [
nor       CC
```

[12] (1) corpus Lancaster; (2) corpus SUSANNE; (3) Penn Treebank; (4) IBM Paris; (5) TOSCA; (6) ENGCG.

knocker	NN1	]N]P]V]Fr]N]
,	,	
was	VBDZ	[V
blistered	VVN	
and	CC	
distained	VVN	V]
.	.	

Exemple 1. Format vertical pour l'annotation grammaticale (extrait de EAGLES 1996).

Le format horizontal ou linéaire présente la phrase analysée en regroupant les constituants dans des parenthèses qui identifient le type de constituant, le tout étant présenté linéairement. Par exemple, dans la phrase suivante issue du corpus IBM Paris (Langé 1994), *ce guide leur permet de se familiariser avec les opérations de réseau local effectuées par les utilisateurs*[13], on procède à l'étiquetage morphosyntaxique et syntaxique de la manière suivante :

[N Ce_DDEMMS guide_NCOMS **N] [V** [P leur_PPCA6MP P] permet_VINIP3
[P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [P avec_PREP
[N les_DARDFP opérations_NCOFP [P de_PREPD [N réseau_NCOMS
[A local_AJQMS A]N]P] [A effectuées_VTRPSFP [P par_PREP
[N les_DARDMP utilisateurs_NCOMP N]P]A]N]P]Vi]P]**V]** .

Exemple 2. Format horizontal (IBM Paris Treebank).

Les crochets identifient les syntagmes : [N …N] pour le syntagme nominal, [V …V] pour le syntagme verbal, etc. Il est possible de rendre l'annotation syntaxique plus explicite graphiquement grâce à l'indentation. Les enchâssements sont indiqués par des retraits de paragraphes. C'est ce que fait le schéma d'annotation TOSCA et ce que permet aussi le corpus IBM Paris.

```
[N Vous_PPSA5MS N]
[V accedez_VINIP5
        [P a_PREPA
                [N cette_DDEMFS session_NCOFS N]
        P]
        [Pv a_PREP31 partir_PREP32 de_PREP33
                [N la_DARDFS fenetre_NCOFS
                        [A Gestionnaire_AJQFS
                                [P de_PREPD
                                        [N taches_NCOFP N]
                                 P]
                         A]
                 N]
         Pv]
V]
```

Exemple 3. Annotation en constituants dans le corpus IBM Paris Treebank.

[13] Dans cet exemple, l'ambiguïté a déjà été levée pour l'étiquetage morphosyntaxique du mot *guide.* Pris tel quel, le mot peut être un nom commun ou un verbe.

Certains corpus permettent une visualisation graphique de l'analyse syntaxique en donnant accès à la structure arborescente de la phrase analysée. C'est le cas par exemple du corpus journalistique français (Abeillé *et al.* 2003).

c.2. Le corpus des verbes de perception sensorielle

Nous définissons dans cette section l'ensemble des annotations que comprend notre corpus. Dans la section consacrée au codage (II.3.4.), nous exposerons comment elles sont insérées dans le corpus. Garside et al. (1997) établissent une différence entre étiquettes (*tags*) et annotations (*labels*). Ceci recouvre une différence dans la manière d'apposer des informations au texte brut. Une étiquette est l'ensemble formé par le mot et les informations qui lui sont ajoutées. Tous les exemples vus jusqu'à présent étaient des étiquettes. Ainsi, « session_NCOFS» est une étiquette morphosyntaxique puisqu'il s'agit d'un seul bloc composé d'une part du mot « session » et d'autre part des informations morphosyntaxiques concernant le mot *session* : « NCOFS » (nom commun féminin singulier). De même [N cette session N] serait une étiquette syntaxique puisqu'elle combine le mot et son annotation syntaxique. Une étiquette est donc le mot avec ses informations. À l'opposé, lorsqu'on parle d'annotation, on fait référence aux seules informations qui ne sont, cette fois, pas attachées au mot auquel elles se rapportent : « NCOFS » seul est une annotation, de même [N] est une annotation. Nous avons choisi pour notre corpus le système de l'annotation plutôt que celui l'étiquetage. Ceci n'empêche pas l'analyse syntaxique. Par conséquent, il n'y a pas de délimitation des constituants dans notre analyse syntaxique, mais simplement une catégorisation (les constituants sont donc nommés au moyen d'annotations).

Par ailleurs, entre l'annotation partielle et l'annotation intégrale (Habert et al. 1997), nous avons opté pour l'annotation partielle (aussi appelée *squelettique*), car seul un sous-ensemble des données est pertinent pour la recherche envisagée. En effet, seuls les constituants principaux sont utiles dans la perspective de notre recherche. L'objectif étant d'étudier la sous-catégorisation des verbes, leur cadre syntaxique et leurs arguments typiques, il est intéressant de savoir que le verbe *entendre* est suivi, par exemple, d'un syntagme prépositionnel, mais il est inutile dans la plupart des cas de connaître la composition exacte de ce syntagme. L'annotation partielle cherche donc à identifier les patrons syntaxiques dans lesquels les verbes étudiés apparaissent.

Dans la phrase suivante, l'annotation squelettique se limitera au verbe (V) et au complément prépositionnel (SPpar) :

Exemple 4. […] les socialistes luxembourgeois ne sont pas entendus par leur président de parti.
(entendre_LaMeuse_031001_23)

L'annotation grammaticale pour le corpus des verbes de perception sensorielle ne présente qu'une seule couche, qui est la catégorisation des syntagmes les plus importants. L'ensemble des annotations utilisées pour cette catégorisation mélange syntaxe et morphosyntaxe. En théorie, il n'existe pas un jeu d'annotations meilleur qu'un autre. En pratique, cela dépend de l'enjeu de l'étiquetage. Mais, surtout, il ne faut pas oublier qu'un problème essentiel demeure : il n'existe ni véritable uniformité ni consensus surtout pour le choix des catégories syntaxiques.

L'important reste d'avoir un ensemble cohérent permettant de transcrire les informations nécessaires à l'analyse linguistique visée. Dans notre corpus, les sigles qui indiquent l'information grammaticale suivent les recommandations du groupe EAGLES (1996) pour les annotations syntaxiques. Les catégories syntaxiques recommandées sont la phrase, la clause (cette catégorie comprend les corrélatives, les relatives, les complétives, etc.), les syntagmes nominal, verbal, adjectival, adverbial et prépositionnel. Notre liste de catégories reprend toutes celles que nous venons d'énumérer, avec comme particularité l'éclatement de la catégorie *clause* en ses différentes sous-catégories. D'autres annotations relevant de la morphosyntaxe ont été ajoutées et c'est une des particularités de l'annotation grammaticale de notre corpus que d'être un mélange entre annotations syntaxiques et morphosyntaxiques. Dans notre cas, l'annotation étant manuelle, le prérequis de l'annotation morphosyntaxique ne se pose pas. Par ailleurs, notre analyse reposant essentiellement sur les types de constituants qui entourent les verbes de perception sensorielle, cette annotation s'avère utile, mais de manière limitée pour les fins d'analyse que nous poursuivons. Par conséquent, notre annotation morphosyntaxique est partielle et ne s'attache qu'à certains phénomènes spécifiques pouvant influencer le sens d'une construction. C'est le cas par exemple de la présence d'un auxiliaire modal (exemple 5), d'une tournure passive (exemple 6), d'un clitique impersonnel (exemple 7) ou encore de l'emploi du gérondif (exemple 8). Par exemple, l'identification de ces éléments dans les phrases suivantes est importante pour le sens attaché à la phrase :

Exemple 5. *Il peut entendre le discours.*
Exemple 6. *Il a été entendu hier.*
Exemple 7. *Il est entendu que ceci est juste.*
Exemple 8. *En entendant cela, il partit.*

Dans les annotations syntaxiques optionnelles proposées par EAGLES (1996), on relève les annotations d'autres constituants tel l'auxiliaire. Nous n'avons pas repris l'indication de l'auxiliaire puisque nous cherchions à annoter seulement les patrons syntaxiques des verbes étudiés. Par contre, le repérage du passif s'est révélé important pour l'analyse. L'annotation de cet auxiliaire-là est donc utile et a été insérée dans la liste des annotations[14]. Les différences entre syntaxe et morphosyntaxe sont indiquées dans le tableau suivant, qui reproduit la liste des annotations grammaticales utilisées :

SIGNIFICATION DES ANNOTATIONS GRAMMATICALES	ANNOTATIONS
1. Catégories syntaxiques (PS)	
Syntagme nominal	SN
Syntagme verbal	SV
Syntagme prépositionnel	SP
Syntagme adjectival	SAdj
Syntagme adverbial	SAdv

[14] *Le policier a entendu le prévenu* obtient le patron [V SN] et la phrase *Le prévenu a été entendu par le policier* obtient le patron [Pass V SPpar].

Phrase	P
Subordonnée	Sub
Comparative	Comp
Corrélative	Corr
Complétive	Compl
Phrase	P
2. Catégories morphosyntaxiques (POS)	
Clitique sujet	CLs
Clitique direct	CLdir
Clitique indirect	CLind
Clitique impersonnel	CLi
Clitique pronominal	CLse
Autres pronoms directs	PROdir
Autres pronoms indirects	PROind
Conjonction de coordination	Coord
Auxiliaire passif	AuxPass
Auxiliaire modal	AuxMod
Passif	Pass

Tableau 2. Annotations grammaticales (POS et PS).

Le corpus comprend d'autres informations syntaxiques et morphosyntaxiques qui permettent d'atteindre un degré de précision essentiel pour l'analyse des données. En effet, grâce aux distinctions supplémentaires apportées par ces annotations, les exemples peuvent être répartis dans des classes aux frontières relativement nettes. Les informations apportées permettent d'indiquer les phénomènes syntaxiques de la négation et de l'incise et la morphosyntaxe du verbe en spécifiant les modes qui ont une incidence sur le sens de la phrase (gérondif, participe présent, impératif).

d. L'annotation sémantique

d.1. Les corpus de référence

Actuellement, on distingue 2 grandes catégories d'annotations sémantiques (Véronis 2000) : l'étiquetage du sens des mots et l'étiquetage des relations dans la phrase et dans le discours. La première catégorie, le choix du sens d'un mot dans un contexte, impose de résoudre les problèmes de polysémie. De manière plus générale, discriminer le sens à attribuer à un mot dans un contexte est une tâche importante en TAL (traduction, recherche d'information, etc.). La seconde catégorie d'annotations sémantiques, c'est-à-dire le marquage des thèmes discursifs comme l'anaphore, est un phénomène très complexe qui en est encore à ses débuts dans les corpus annotés.

La standardisation de ces annotations ne peut avoir lieu que s'il y a un accord sur l'analyse linguistique, or ce n'est pas le cas. Si l'on se base sur le recensement des mots et des différents sens qui leur sont associés tel qu'il est présenté dans les dictionnaires, il faut tenir compte du fait que cette liste des sens pour un mot donné varie d'un dictionnaire à l'autre, de même que leur description. Par ailleurs, l'étiquetage automatique du sens des mots est, pour

l'instant, « un thème de recherche plus qu'une technique pouvant être appliquée [automatiquement] aux corpus. » (Véronis 2000, p. 120) Des tentatives se font pour l'anglais (Miller et al. 1993, Ng & Lee 1996, Weibe et al. 1997), mais il n'existe encore aucune ressource pour le français. En d'autres termes, l'étiquetage / annotation ne peut se faire que manuellement.

d.2. Le corpus des verbes de perception sensorielle

Dans notre corpus, l'annotation sémantique est manuelle et c'est la raison pour laquelle elle est faisable. L'effort important que l'on voit actuellement pour développer des ontologies –par exemple Wordnet (Miller *et al.* 1993, Fellbaum 1998) ou EuroWordnet (Vossen 1998)– ne permet pas d'apporter un éclairage nouveau sur les critères distributionnels qui sont à la base du choix d'un sens dans un contexte particulier (Véronis 2000). Or, c'est justement l'étude que nous faisons. L'étiquetage sémantique de notre corpus se base donc sur les listes de sens des dictionnaires. Notre annotation indique le sens que prend le verbe de perception analysé dans la phrase dans laquelle il apparaît. Une première version de la liste des sens que peuvent prendre chacun des verbes de perception étudiés a été établie lors de l'analyse lexicale (section III. 1.2.) et elle a été révisée au cours de l'analyse de corpus.

D'autres informations sémantiques sont également insérées dans notre corpus. Elles ont une certaine importance dans le choix du sens donné à un bon nombre de phrases. Lorsque ces informations s'avèrent utiles, ou tout le moins complémentaires aux informations syntaxiques et morphosyntaxiques, elles sont insérées dans le corpus. Par exemple, lorsque le verbe de perception sensorielle est suivi d'une complétive, notamment d'une complétive infinitive, il est utile de savoir quel verbe est employé dans la complétive. En effet, il existe une différence de sens au niveau du verbe *entendre* entre *entendre parler* et *entendre accomplir une tâche*[15].

Également, on retrouve très souvent dans le corpus l'identification du type de syntagme nominal complément ou sujet du verbe de perception : syntagme nominal indiquant un son, la source d'un son, etc. Ces catégories sont issues de l'analyse cognitive reposant sur la physiologie de la perception (section III. 2.2.).

Maintenant que les types d'annotations apposées à notre corpus ont été définis, nous présentons dans la section suivante le codage de ces annotations.

II. 3.4. Codage et normalisation des ressources linguistiques

a. Problématique et objectifs

Une fois le corpus construit et les choix d'étiquettes faits, il faut pouvoir encoder ce corpus de manière à ce qu'il puisse être utilisé pour l'analyse linguistique et pour des applications diverses.

[15] La différence de sens s'accompagne d'une autre différence : le sujet de l'infinitif n'est pas le même dans ces deux phrasesé.

Les règles émises par la TEI (*Text Encoding Initiative*), version P4 (2001), ont pour but d'uniformiser l'encodage des données textuelles. Les recommandations qui y sont faites proposent de créer des documents qui sont conformes soit au format SGML (seul format recommandé dans la version P3 [1999]) soit au format XML. Dans les deux cas, il s'agit de langages de balisage (*markup language)*. Dans le corpus que nous avons construit, nous suivons la recommandation pour le choix de l'encodage en XML. Par contre, il ne s'agit pas d'un texte exactement tel que l'envisage le guide de la TEI. Sa spécificité de corpus d'échantillons ne peut reproduire les balises proposées. Nous avons opté pour un encodage différent, mieux adapté à notre type de corpus et à nos objectifs d'analyse. Toujours est-il que le choix de XML permet d'avoir un format d'échange standard. Quant au choix des étiquettes (section III. 3.3.), il se base sur les règles émises par le groupe EAGLES, qui sont beaucoup plus détaillées, plutôt que sur les règles émises par la TEI pour l'annotation linguistique (http://www.tei-c.org/P4X/AI.html#AILA – section 15.4).

b. Documents XML

b.1. Particularités

Les avantages du format XML, comme de tout langage de balisage, sont l'emphase sur la description, le concept de descriptions des catégories et l'indépendance par rapport à la plate-forme et aux logiciels. Tout d'abord, les codes XML permettent de catégoriser les différentes parties du document, sans aucune indication procédurale. C'est ainsi qu'en XML la description est séparée de la procédure à suivre. Grâce à ce marquage descriptif plutôt que procédural, le document peut être traité seulement en fonction des parties intéressantes pour la tâche à faire. Par ailleurs, plusieurs programmes peuvent être appliqués au même document. Un autre élément clé du format XML est la description des catégories (soit dans une DTD [*Document Type Definition*] soit dans un schéma) qui est attaché à chaque document XML. Il définit formellement les parties qu'un document doit posséder, ainsi que leur structure. La DTD est le document type le plus populaire. Il s'agit en fait un modèle issu de SGML. Sa syntaxe est donc différente de celle d'un document XML. De plus, il n'est pas possible de spécifier ni les patrons pour les données dans les éléments ni les attributs. Un schéma XML (parfois appelé XSchéma) définit la structure d'un document XML, tout en respectant la syntaxe XML et en donnant davantage de contrôle sur les types de données. Les éléments, tout comme les attributs, peuvent avoir des listes de valeurs possibles (Ray 2003).

Enfin, un document encodé en XML peut être utilisé par des plates-formes différentes et des logiciels différents. De fait, tous les documents XML utilisent le même encodage de caractères, défini par un standard international. En comparaison avec HTML, un autre langage de balisage, XML présente trois caractéristiques intéressantes. La première est que XML ne contient pas un nombre fixe d'étiquettes : c'est un langage extensible. La deuxième caractéristique est que les documents XML doivent être construits selon une syntaxe établie dans un document type qui sert à leur validation. Enfin, XML est un langage dont l'emphase porte non pas sur la présentation des données, mais sur leur position dans la structure arborescente.

b.2. Les corpus de référence

Les corpus les plus récents sont conformes au format XML. C'est le cas du projet TIGER Treebank pour l'allemand (Brants & Hansen 2002). Il s'agit d'un corpus de journaux allemands de 55 000 phrases annotées (POS et syntaxe) de manière semi-automatique. L'analyse syntaxique relève du formalisme LFG (*Lexical Functional Grammar*).

Un autre projet de corpus encode ses données en XML. Il s'agit du programme d'archivage de données du groupe LACITO (LAngues et CIvilisations à Tradition Orale). Le programme vise à archiver des documents qui associent une transcription textuelle à des enregistrements d'énoncés dans des langues sans écriture. Dans ce cas, le but visé n'est pas de produire une analyse syntaxique du corpus, ni une annotation en catégories, mais de donner une transcription phonétique, une traduction et des informations sur des phénomènes utiles à décrire.

Pour le français, on relèvera le projet de corpus arboré de l'équipe de Abeillé (Abeillé & Clément 1999, Abeillé *et al.* 2003). Ce corpus suit les règles émises par la TEI, tout en les adaptant, et présente les annotations grammaticales en format XML. Voici un exemple tiré de ce corpus :

```
<SENT><PP>Au cours de:P
        <NP> la:Dfs conf´erence de presse:NC-fs
                <Srel> <NP>:SUJ qui:PROR-3fs </NP>
                        <VN> a:VP-3s clos:VK-ms </VN>
                        <NP> cette:D-fs rencontre:NC-fs </NP>
                </Srel>
        </NP> </PP> ,:PONCT
<NP>le:D-ms premier ministre:NC-ms<AP>est-allemand:A-ms</AP></NP>
<VN>est:VP-3s revenu:VK-ms</VN>
<PP> sur:P <NP>les:D-mp incidents:NC-mp
        <PP> de:P <NP>lundi:NC-ms soir:NC-ms</NP></PP>
</NP> </PP>
</SENT>
```

Exemple 9. Un extrait du corpus arboré de l'équipe de Abeillé (Abeillé et al. 2003).

D'autres projets voient le jour. Par exemple, l'étude de Estival & Nicholas (1999) qui suit les règles de la TEI pour encoder la présentation d'un texte français du début du XIVe siècle (Jehan de Joinville, *La vie de Saint-Louis*) dans le but d'une étude diachronique. L'étiquetage syntaxique se fait ensuite en XML et adapte les règles d'annotation syntaxique émises par la TEI (Sperberg-McQueen & Burnard 1995). Selon les règles de Sperberg-McQueen & Burnard émises en 2001, voici ce que donnerait l'encodage syntaxique d'un début de phrase *the victim's friends told…* :

```
<s type="sentence">
  <phr ana="n">
    <phr ana="g">
      <w ana="at">The</w>
      <w ana="nn1">victim</w>
      <m ana="gen">'s</m>
    </phr>
```

```
    <w ana="nn2">friends</w>
  </phr>
  <phr ana="v">
    <w ana="vvd">told</w>
    ...
</s>
```

Exemple 10. Un extrait selon les règles d'annotation linguistique émises par la TEI (2001).

b.3. Le corpus des verbes de perception sensorielle

Comme il a été dit plus haut, un document XML est associé à un autre document qui définit sa structure[16]. Notre corpus, un document XML, est associé à un schéma qui définit la structure des étiquettes du document XML et les contraintes sur l'information qui peut y figurer[17].

Le corpus établi pour les verbes de perception est divisé en quatre parties puisqu'il y a un corpus dédié à chacun des verbes de perception : *voir, entendre, goûter* et *sentir*. Les quatre corpus seront encodés en XML et la structure de chacun de ces documents sera identique. Elle est définie au moyen d'un schéma. La seule différence qui existe entre les schémas des corpus est leur racine, qui est le verbe auquel est dédié le corpus. Par contre, l'organisation des attributs et de leurs valeurs qui viennent à la suite de la racine est identique. Voici l'organigramme de ce schéma avec *entendre* en racine :

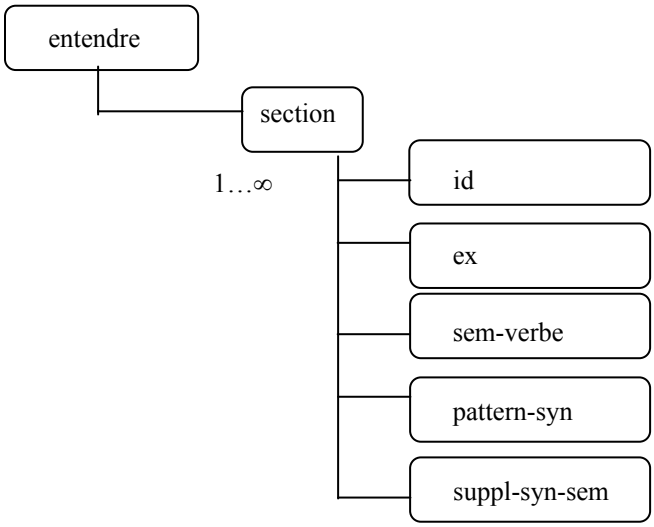

Figure 2. La structure du schéma XML pour l'encodage du corpus des verbes de perception sensorielle.

[16] Sans pour autant que ce document associé soit une obligation.

[17] Schéma et ontologie diffèrent. Une ontologie est une conceptualisation explicite (Fensel et al. 2003), c'est une théorie du domaine et non une structure de données, telle que le sont les langages de schémas (schéma XML, OIL [Ontology Inference Layer], schéma RDF). Les schémas XML ne sont d'ailleurs pas faits pour la modélisation d'ontologies puisqu'ils décrivent la structure obligatoire d'un document. Le schéma, tout comme la DTD, contraint la structure syntaxique des étiquettes XML. Le schéma XML donne un balisage sémantique qui se révèle utile pour le « web sémantique » (Fensel et al. 2003).

Le verbe auquel est dédié le corpus est la racine du schéma : *entendre*, par exemple. Sous cette racine apparaît un nombre non limité de sections. Le document se divise ainsi en sections qui contiennent chacune une suite d'éléments : *id, ex, sem-verbe, pattern-syn* et *suppl-syn-sem*. L'étiquette <id> constitue l'identification de la section. Les références utilisées identifient chaque phrase du corpus en combinant des informations de plusieurs types. D'abord, le verbe dont il est question, le sous-corpus duquel l'exemple est tiré, la datation de l'exemple et la numérotation de l'exemple au sein du sous-corpus. Par exemple, entendre_LaMeuse_021001_04.

L'élément suivant dans une section est l'élément <ex>. Il contient une seule concordance à la fois. Dans le cas du *Robert Électronique,* l'extrait est la citation au complet, telle qu'elle est donnée dans le dictionnaire. Dans les concordances issues du journal *La Meuse*, le concordancier n'envoie qu'un extrait comprenant un certain nombre de caractères avant et après le terme demandé. Par conséquent, les extraits sont souvent tronqués. Cela dit, la sélection des extraits se faisant sur la base d'une fenêtre de 80 caractères avant et après le verbe, les extraits ne sont tronqués qu'en dehors de la proposition dans laquelle apparaît le verbe étudié. L'analyse n'est donc pas gênée par cet aspect. Nous avons inséré des crochets à l'intérieur de l'extrait de manière à délimiter, à l'intérieur de l'extrait obtenu par le concordancier, la proposition qui est soumise à l'analyse (exemple 11).

Exemple 11. tiré la sonnette d'alarme. Un appel au secours [auquel la Ville de Verviers entend manifestement répondre.] Pour l'école des jeunes Jusqu'à présent, la

L'étiquette <sem-verbe> indique le sens que le verbe analysé a dans l'extrait. L'étiquette suivante, <pattern-syn>, contient les informations sur le patron syntaxique auquel l'extrait analysé appartient. Par exemple, on a dans la phrase ci-dessus [PROInd' SN *V* SAdv [V']]. Le sigle *V* identifie le verbe de perception auquel le corpus est consacré. Les apostrophes relèvent les relations de dépendance éloignées. Dans cet exemple, le pronom relatif (PROInd) est attaché au verbe infinitif de la complétive (V, *répondre*). Enfin, l'étiquette <suppl-syn-sem> comprend des informations supplémentaires diverses, qui relèveront, selon les exemples, de la syntaxe ou de la sémantique. On pourra y trouver le verbe utilisé dans la complétive, la présence d'un adverbe de qualité, etc. Lorsqu'on assemble tous ces éléments, une section du corpus se présente comme suit :

```
<section id="entendre_LaMeuse_191001_11">
        <ex>tiré la sonnette d'alarme. Un appel au secours [auquel la Ville de Verviers entend manifestement
répondre.]Pour l'école des jeunes Jusqu'à présent, la</ex>
        <sem-verbe> volonté</sem-verbe>
        <pattern-syn>PROInd' *V* [V']</pattern-syn>
        <suppl-syn-sem>SAdv (manifestement) + répondre</suppl-syn-sem>
          </section>
```

Exemple 12. Une section dans le corpus des verbes de perception sensorielle.

La particularité de l'encodage de notre corpus est d'avoir inséré les informations additionnelles, non pas sous forme d'étiquettes (donc d'annotations accrochées à chaque mot ou groupe de mots), mais dans des couches dédiées à

des informations relevant de différents domaines. Ainsi, l'information est stratifiée. Les annotations grammaticales apparaissent dans les couches réservées à ce type d'informations (pattern-syn et suppl-syn-sem); les informations sémantiques, dans les couches réservées à cet effet (sem-verbe et suppl-syn-sem). Il est ainsi possible de séparer le texte brut des informations qui lui sont associées. Le classement des exemples peut se faire sur l'une ou l'autre des catégories présentées.

II. 4. GESTION DES RESSOURCES LINGUISTIQUES

L'étape suivant l'encodage du corpus des verbes de perception sensorielle est sa présentation sous une forme lisible pour que ce corpus puisse être consulté. Une feuille de style (XSLT ou XSL-FO) sera attachée au document XML. Les différentes informations contenues dans le corpus seront accessibles sur requête : le choix sera donné de rechercher des citations selon le type de corpus, selon le sens du verbe ou le patron syntaxique dans lequel le verbe peut être employé. Il est vraisemblable que les choix données seront plutôt statiques, vu les autres aspects que le projet se doit de développer.

III. ANALYSE LINGUISTIQUE

III. 1. LES RESTRICTIONS SÉLECTIONNELLES ET LE SENS DU VERBE

III. 1.1. Problématique et objectifs

L'étude diachronique permet de mieux déceler les champs sémantiques qui se partagent l'espace sémantique d'un verbe. Une fois ces champs établis, l'étude des acceptions en coupe synchronique est plus aisée. La consultation des dictionnaires et des ressources linguistiques (lors de la constitution du corpus) a permis, conjointement, d'établir une première version de la liste des acceptions cette fois plus raffinée que ce que l'analyse diachronique et la consultation des dictionnaires avaient permis d'obtenir (section II. 2.).

L'analyse du corpus construit va permettre de décrire les sens que prend chaque verbe de perception dans les emplois issus du corpus littéraire d'une part et du corpus neutre d'autre part. Il s'agit de découvrir des régularités syntaxiques et sémantiques qui permettraient de guider le choix du sens à donner au verbe selon les contextes. Ces régularités une fois découvertes seront encodées dans le lexique computationnel (section V).

III. 1.2. Un exemple d'analyse : le cas du verbe *entendre*

Le but de cette étude étant de faire correspondre des significations avec des contraintes syntaxiques et des restrictions sélectionnelles, il faut disposer initialement de la liste des sens des verbes étudiés. Prenons le cas de *entendre.* Les notices de trois dictionnaires de référence pour l'usage contemporain (*Le Robert électronique* 1994, *Le grand usuel Larousse* 1997 et le *Trésor de la langue française* 1971 – 1994) ont été compilées. Par ailleurs, le travail sur le corpus a permis de mettre en lumière certaines nuances qui étaient absentes des dictionnaires consultés. La

liste des sens ainsi établie comprend 15 entrées qui se répartissent en deux domaines sémantiques : le domaine sensoriel, qui relève donc de l'audition (7 entrées), et le domaine mental (8 entrées). Ces sens ont été intégrés aux annotations du corpus et ont ainsi été mis en correspondance avec les constructions syntaxiques.

I. DOMAINE SENSORIEL	**Exemples**
1. Compétence auditive	Je n'entends plus rien de l'oreille droite.
2. Perception auditive	J'entends le son de la cloche.
3. Attention auditive (écoute)	J'entends une conférence sur le verbe *entendre.*
4. Écoute officielle	Les jurés entendent les témoins.
5. Écoute favorable	Les revendications ont été entendues.
6. Mémoire auditive	On entend encore ses paroles.
7. Transgression sensorielle	J'entends les couleurs.

II. DOMAINE MENTAL	**Exemples**
8. Accepter	Étant entendu qu'il faut trouver une solution, …
9. Juger	À entendre ses paroles, tout était perdu.
10. Perception d'une situation	La multinationale ne l'entendait pas ainsi.
11. Volonté	L'entraîneur entend conserver ce joueur.
12. Relation - accord	Nous devrons nous entendre sur ce point.
13. Relation réciproque	Ils sont faits pour s'entendre.
14. Compréhension	Je me fais entendre de mes interlocuteurs.
15. Compréhension - habileté	Il s'y entend pour troubler les esprits.

Tableau 3. La liste provisoire des sens établis pour le verbe *entendre*.

L'annotation grammaticale partielle du corpus a permis de regrouper les emplois en un nombre restreint de patrons syntaxiques. Ces patrons syntaxiques ont été établis pour classer les emplois du verbe *entendre* tel qu'il apparaît utilisé dans le corpus mis sur pied. Ces patrons encodent la sous-catégorisation syntaxique du verbe et, par conséquent, s'attachent uniquement à la complémentation essentielle du verbe. Ce qu'il faut savoir, c'est si le verbe est employé avec un syntagme nominal ou s'il peut être utilisé avec un syntagme prépositionnel essentiel, avec une complétive infinitive, etc. C'est ce que les patrons syntaxiques permettent de découvrir. Pris tels quels, les types de constructions syntaxiques sont nombreux. Nous en avons relevé une bonne soixantaine. L'analyse a permis de constater que ces types se regroupaient en 9 patrons. Ainsi regroupés, ils permettent d'étudier le comportement syntaxico-sémantique du verbe.

Patrons syntaxiques	**Exemples**	**Référence des exemples**
[*V* SN]	[Nous n'entendons pas la musique mais les basses.]	entendre_LaMeuse_241001_09
[*V* [V]]	[Mais les anciens combattants entendent demander à ce que l'éventuel nouveau sanglier soit en pierre et scellé]	entendre_LaMeuse_031101_14
[*V* [x V]]	[Il entendait l'accomplir par tous les moyens dont il disposait.]	entendre_GrandRobert_cit20_492
[*V* Compl]	[pour laisser entendre que le salaire était variable.]	entendre_GrandRobert_cit20_183
[*V* P]	["Des films de cette qualité, on en redemande" entendait -on à la sortie de la projection.]	entendre_LaMeuse_131101_09
[Pass *V*]	[L'affaire est entendue.]	entendre_GrandRobert_cit20_053

[CLse *V*]	[Dès l'entame du 5ème set, les supporters se font entendre.]	entendre_LaMeuse_291101_13
[*V*]	[qui nous permettent de voir et d'entendre très au-delà de ce qu'auraient pu rêver les plus audacieux futurologues]	entendre_GrandRobert_cit20_311
[CLind *V*]	[qu'il n'y *entendait rien,]	entendre_GrandRobert_cit20_136

Tableau 4. La liste des patrons syntaxiques du verbe *entendre* dans le corpus analysé.

La fréquence des patrons syntaxiques est utile à analyser dans la mesure où elle peut déboucher sur de fortes différences d'utilisation d'un patron à l'autre. Une telle information s'avère intéressante à encoder dans des systèmes de TAL. Dans le cas du verbe *entendre,* les neuf patrons syntaxiques qui ont été relevés se répartissent très variablement au sein des deux corpus. Tout d'abord, de manière générale, on constate que la construction de type [*V* SN] est, de loin, la plus fréquente. Les autres constructions que l'on retrouve le plus souvent sont [*V* [V]], ensuite [Pass *V*] et enfin [CLse *V*]. Les constructions [*V* [x V]] et [*V* Compl] sont beaucoup moins représentées. Enfin, les constructions [*V*], [*V* P] et [CLind *V*] sont sous-représentées.

Patron syntaxique	La Meuse		Le Robert		Corpus général	
V SN	552	40,59%	377	52,36%	929	44,66%
V [V]	316	23,24%	158	21,94%	474	22,79%
Pass *V*	182	13,38%	35	4,86%	217	10,43%
CLse *V*	127	9,34%	55	7,64%	182	8,75%
V […V]	68	5,00%	40	5,56%	108	5,19%
V Compl	73	5,37%	30	4,17%	103	4,95%
V	18	1,32%	20	2,78%	38	1,83%
V P	24	1,76%	4	0,56%	28	1,35%
CLind *V*	0	0,00%	1	0,14%	1	0,05%
Total	1360	100,00%	720	100,00%	**2080**	**100 %**

Tableau 5. La fréquence des patrons syntaxiques du verbe *entendre* dans le corpus analysé.

Des différences apparaissent entre le corpus journalistique et le corpus littéraire en ce qui concerne la construction [*V* SN], qui est encore davantage représentée dans le corpus littéraire avec 52,36 % (comparé à 40,59 % dans le corpus journalistique). À l'inverse, le corpus issu du journal *La Meuse* utilise presque trois fois plus la construction [Pass *V*].

De manière à guider le choix du sens à donner à une construction, des critères syntaxiques sont bien sûr essentiels, mais il faut aussi tenir compte de ce que l'on nomme *les préférences sélectionnelles* ou *les restrictions sélectionnelles* (Manning & Schütze 2000). Il s'agit des types sémantiques qui caractérisent les arguments d'un verbe. L'exemple le plus souvent cité est le cas du verbe *manger*, qui se construit avec des arguments relevant du domaine de la nourriture. On parle de *préférences*, parce que ce ne sont là nullement des règles. Pensons aux emplois métaphoriques qu'un verbe peut présenter.

Nous avons intégré à notre corpus des annotations qui identifient l'information sélectionnelle des arguments du verbe *entendre.* De nombreux travaux se sont basés sur WordNet ou GermaNet pour identifier les préférences sélectionnelles (citons seulement Resnik 1997, Clark & Weir 2002). La particularité des informations que nous avons insérées est d'être basées sur la physiologie de la perception auditive. Ce type de perception distingue trois modes auditifs (Mann & Liberman 1983; Perez 1994) : environnemental, phonétique et musical. Ces distinctions se sont avérées utiles pour notre analyse.

Nous ne donnerons ici que l'analyse pour le patron syntaxique le plus représenté : le verbe *entendre* suivi d'un syntagme nominal. Cette construction est la plus fréquente, avec 44,66 % d'exemples du corpus. C'est en même temps une construction très polysémique puisqu'elle comptabilise 12 sens différents dans les deux corpus. Le tableau ci-dessous présente la répartition des sens correspondant à la construction [*V* SN] d'abord dans le corpus La Meuse, ensuite dans le corpus du Robert électronique et enfin, dans les deux corpus pris conjointement. Bien évidemment, la répartition entre ces sens est très inégale. Sur l'ensemble des deux corpus, tous les sens relevant du domaine de l'audition (les 7 premiers sens du tableau ci-dessous) atteignent une représentation de 78,59%. Et, à l'intérieur de ce domaine, c'est la perception auditive qui correspond le plus souvent à la construction [*V* SN]. Le corpus littéraire fait d'ailleurs un usage très important de cette construction avec le sens de la perception auditive (75,33 %). Le corpus journalistique montre le même attrait pour la correspondance entre la construction [*V* SN] et le sens de la perception auditive (47,28 %), mais il montre surtout davantage de répartition des sens du domaine auditif en faisant un usage, plus restreint il est vrai, des sens d'écoute - attention (17,39 %) et d'écoute officielle (11,78 %).

Sens de *entendre*	**La Meuse**		**Le Robert**		**Corpus général**	
perception auditive	261	47,28%	284	75,33%	545	58,67%
écoute - attention	96	17,39%	13	3,45%	109	11,73%
écoute officielle	65	11,78%	2	0,53%	67	7,21%
écoute favorable	4	0,72%	0	0,00%	4	0,43%
compétence auditive	1	0,18%	1	0,27%	2	0,22%
transgression modale	1	0,18%	1	0,27%	2	0,22%
mémoire auditive	0	0,00%	1	0,27%	1	0,11%
comprendre	53	9,60%	61	16,18%	114	12,27%
accepter	21	3,80%	10	2,65%	31	3,34%
perception - situation	29	5,25%	2	0,53%	31	3,34%
juger	11	1,99%	2	0,53%	13	1,40%
vouloir	10	1,81%	0	0,00%	10	1,08%
total	552	100,00%	377	100,00%	929	100,00%

Tableau 6. La fréquence des patrons syntaxiques du verbe *entendre* dans le corpus analysé.

En ce qui concerne les sens relevant du domaine mental, la compréhension correspond le plus souvent à la construction [*V* SN]. Ici encore, le corpus littéraire présente une forte polarisation de la construction étudiée envers

un sens particulier, celui de la compréhension. Le corpus journalistique utilise lui aussi souvent la construction transitive directe pour exprimer la compréhension, mais il montre plus de variété en exprimant les sens de *accepter,* de *perception – situation,* de *juger* et de *vouloir* au moyen de cette même construction.

Il nous est impossible de présenter ici toutes les distinctions qui concourent à déterminer le sens associé à la construction [*V* SN]. Par contre, nous insisterons sur les éléments les plus saillants que notre analyse a permis de faire ressortir. En premier lieu, la caractérisation sémantique du SN qui est complément du verbe *entendre* permet de définir les grands choix de sens opérés. Un SN de type *musique* ou *source_musique* (7,53% du corpus général, avec la construction [*V* SN]) ne permet, dans le corpus étudié, que deux interprétations relevant d'ailleurs du domaine de l'audition : la perception auditive ou l'écoute. Un SN de type *parole* ou *source_parole* (39,07% du corpus général) peut prendre de nombreux sens, mais une très nette préférence va vers la perception auditive (60,88% des SN de ce type prennent ce sens). Les autres sens associés à cette construction sont l'écoute – attention, l'écoute officielle et, dans les citations littéraires, le sens de comprendre. On remarque également que, avec le sens de la perception auditive, le corpus *La Meuse* préfère vraiment utiliser un SN_parole plutôt qu'un SN_source_parole, alors que cette préférence, si elle existe clairement dans le corpus littéraire, y est moins forte. Un SN de type *son* ou *source_son* (26,91% du corpus général) démontre une préférence importante pour le sens de la perception auditive (98,4% des SN de ce type prennent ce sens). Les quelques rares autres cas relèvent de toute façon du domaine général de l'audition. Il est intéressant de souligner un dernier point dans l'analyse de la construction [*V* SN] : il s'agit de l'expression de la compréhension. Dans ce cas-ci, l'analyse repose tantôt sur des critères de catégorisation sémantique du complément du verbe, tantôt sur des spécificités de la construction syntaxique. Les deux types de corpus montrent une préférence pour la construction « syntagme prépositionnel (*par*) – verbe – SN ». Voici un exemple :

> Exemple 13. llissement de la population et l'augmentation de l'obésité. [Par diabète, on entend souvent "diabète sucré", qui se définit par un taux de sucre dans le sang] (entendre_LaMeuse_201001_33)

Dans le corpus neutre, 41,67% des emplois de *entendre* au sens de *comprendre* avec le patron général [*V* SN] utilisent cette construction plus spécifique. Le chiffre monte à 55,76% dans le corpus littéraire. Ce dernier utilise également la construction *V* SN avec un SN de type *parole* ou *hors_classe* pour signifier la compréhension. Ce procédé est délaissé par le registre neutre au profit d'un autre procédé (syntaxique) : l'utilisation de l'impératif (exemple 14).

> Exemple 14. oline soit bien mise» dit-elle au directeur de l'établissement. Caroline? [Entendez sa perruque!] Des petites phrases en wallon, un dialogue quotidien, (entendre_LaMeuse_071101_09)

Cette analyse des préférences sélectionnelles du verbe *entendre* a permis de constater que les restrictions sélectionnelles portant sur les arguments du verbe sont importantes pour guider le sens à donner au verbe dans ce contexte. Nous comptons encoder les résultats de cette analyse dans le lexique computationnel (section V). Nous

avons également pu constater que l'analyse des restrictions sélectionnelles en termes de physiologie de la perception (auditive, dans ce cas) autorise un classement qui repose sur la cognition et qui donne des résultats utiles.

III. 2. SÉMANTIQUE

III. 2.1. Problématique et objectifs

Il existe une assez grande diversité des théories sémantiques pour étudier les langues naturelles. Elles se différencient par les notions qu'elles placent au centre de l'espace sémantique qu'elles définissent. On relèvera ici trois grands axes dans les choix de ces notions.

Une première catégorie de théories postule une sémantique s'appuyant sur la logique. Ces théories caractérisent les principes d'une traduction de la langue naturelle (expressions linguistiques) dans une langue formelle (expressions logiques) et la représentation des expressions linguistiques dans le système formel. Ainsi, des entités (par exemple, *sophie*) vérifient des relations (par exemple, *suivre*) : $\exists$ x(suivre(x,sophie) $\wedge$ cours(x)). La notion de vérité est au centre des théories logiques puisqu'il s'agit toujours de déterminer la vérité ou la fausseté d'une proposition. En effet, connaître sa valeur de vérité, c'est connaître sa signification.

La sémantique des situations (Barwise & Perry 1983) s'est développée sur la base d'une remise en cause fondamentale des entités constitutives de l'univers d'interprétation des langues naturelles, telles que postulées par la majorité des approches logiques de la sémantique. En cela, cette théorie constitue le deuxième axe des théories sémantiques prises en compte ici : la sémantique de Barwise et Perry centre son espace sémantique sur la nature informationnelle des expressions linguistiques. Les entités premières de cette théorie sont les situations et les entités d'information correspondant à ces situations. De plus, au lieu d'associer à une proposition une valeur de vérité unique, la théorie des situations permet qu'une même expression utilisée par plusieurs locuteurs puisse décrire différentes situations. La signification linguistique d'une expression est donc une relation entre des situations d'énonciation et des entités individuelles. La représentation de la situation *Paul rit* <{rire, Paul}> présente un état de fait. Cette sémantique a pour objet non plus le traitement de la langue par le locuteur-auditeur, mais les relations entre les expressions de la langue et l'environnement extérieur de l'agent qui traite ces expressions. Ces entités et leurs relations continuent de faire l'objet d'une théorisation mathématique puisque l'objectif est toujours de développer une théorie sémantique mathématisée.

Le troisième axe qui divise les théories sémantiques est celui des approches cognitives. Elles conçoivent l'activité langagière comme dépendant directement d'un faisceau de facteurs : physiques, biologiques, psychologiques, sociaux, culturels. Le langage n'est donc pas un système autonome et son étude s'insère dans le cadre des sciences cognitives. Ces travaux cherchent à établir les liens qui existent entre la langue et les autres facultés humaines. Ce postulat cognitif place les approches cognitives en opposition diamétrale avec le courant générativiste, qui décrit « la compétence linguistique comme une capacité isolée ayant une évolution propre et indépendante des autres

capacités humaines. » (Chambreuil 1998, p. 343) Par ailleurs, la place centrale accordée à la sémantique dans l'ensemble de ces théories les rend particulières. En effet, même si le courant générativiste intègre de plus en plus de notions sémantiques (Chomsky 1995), il n'en demeure pas moins un courant dans lequel la syntaxe reste primordiale.

Notre objectif est d'étudier les verbes de perception sensorielle dans le cadre de la sémantique cognitive. Tout d'abord, bien sûr, parce que nous adhérons aux principes fondamentaux que pose ce cadre théorique. Il nous semble que faire abstraction de l'expression des facultés humaines au sein du langage revient à voiler une grande partie du fonctionnement langagier. De plus, notre sujet d'étude ne pourrait mieux se prêter à une analyse tenant compte des facultés humaines. Dans la section suivante, nous exposerons plus longuement les positions prises par la linguistique cognitive.

Le courant de la linguistique cognitive doit être considéré comme un paradigme d'analyse plutôt que comme une seule théorie. Les différentes approches qui utilisent le paradigme cognitif se caractérisent toutes par leur emphase sur la relation entre le langage et les autres facultés cognitives. Ces divers cadres théoriques ne s'opposent pas les uns aux autres, ils se différencient plutôt par le type de phénomènes qu'ils analysent.

Le paradigme cognitif a effectué un virage dans la conception de la sémantique (notamment avec Miller & Johnson-Laird 1976). Chaque étiquette donnée à un mot est associée à un concept lexical. Pour la plupart des mots, un concept lexical est composé de deux parties : une partie de définition qui relève du schéma perceptif et une partie qui consiste en la connaissance associée au mot, avec entre autres les relations entre ce concept et les autres concepts. Dans un tel cadre théorique, la sémantique s'insère dans un système de connaissances plus large et englobe les connaissances encyclopédiques, qui sont une partie de la cognition humaine (mémoire conceptuelle). La définition des concepts suit les conclusions des recherches en psychologie (notamment Rosch 1975) en introduisant la notion de prototype avec tout ce qu'elle implique (degré de représentativité, caractéristiques essentielles, relativité des définitions en extension et en intension). Ainsi, en linguistique cognitive, la connaissance d'une entité est non seulement vaste, mais cette connaissance présente également plusieurs facettes. C'est ce que recouvre le terme de *domaine cognitif*. Un terme lexical fait référence à un ensemble de domaines cognitifs, mais l'accès à ces domaines ne se fait pas au hasard. Premièrement, le nombre de domaines lexicaux invoqués est limité et deuxièmement ces domaines sont invoqués dans un certain ordre. Chacun de ces domaines reçoit ainsi un degré de centralité (Lakoff 1987, Langacker 2000). Dans une telle vision de la sémantique, la recherche s'intéresse à la structure interne des concepts en ce qu'elle présente comme changements (polysémie) et comme prototypes (noyaux de sens). Il s'agit donc d'une théorie conceptuelle de la signification selon un point de vue psychologique.

C'est ainsi que de nombreux travaux en linguistique cognitive sont dédiés à la métaphore. Au lieu de la concevoir comme un outil littéraire, ces théories la caractérisent comme un phénomène conceptuel. La métaphore est un déplacement d'un domaine source vers un domaine cible. Selon Lakoff & Johnson (1980), la correspondance entre

les deux domaines préserve toujours une structure schématique imagée. Il y a donc des propriétés configurationnelles qui sont imposées d'un domaine à l'autre et qui permettent la projection de l'un vers l'autre.

La conception la plus fondamentale que l'on retrouve sous des formes diverses dans plusieurs théories cognitives est la notion de schémas imagés (Fauconnier 1985; Lakoff 1987; Langacker 1987, 1991). Ce sont des conceptions abstraites qui trouvent leur origine dans l'expérience physique des êtres humains et qui permettent de structurer leur monde mental. Pour Langacker (2000), ces conceptions sont innées tandis que pour Lakoff (1990) elles sont dérivées de l'expérience. L'utilisation des schémas imagés prend des aspects différents selon les approches. Ainsi, Fauconnier (1985) étudie les relations qui sont établies entre les éléments des modèles mentaux (espaces mentaux) construits par tout locuteur. Ces relations sont à la base de phénomènes sémantiques comme les ambiguïtés de portée. Langacker (2000) insiste sur l'aspect configurationnel des schémas imagés, parmi lesquels on trouve essentiellement les notions de *source-chemin-but, contenant-contenu, centre–périphérie*, etc. Il relève aussi la capacité à établir des relations, à comparer, à faire des abstractions. Pour Langacker, l'être humain possède des habiletés cognitives qui s'intègrent dans la sémantique et plus généralement dans la structure de la langue. La sémantique et la grammaire sont d'ailleurs indissociables puisque la grammaire est la structuration de contenus conceptuels.

Talmy (1988, 2000) s'intéresse quant à lui essentiellement aux relations entre les facultés cognitives générales et le langage. Son approche insiste sur les processus qui permettent d'organiser le contenu conceptuel dans la langue. En d'autres termes, il s'agit d'étudier comment le langage structure le contenu conceptuel. Ainsi, quelle est l'organisation de catégories telles que l'espace, le temps, le déplacement, etc.? Il étudie parallèlement les propriétés formelles de la langue et les relations avec les structures cognitives qui relèvent d'une approche psychologique.

Talmy (2000) relève un certain nombre de systèmes schématiques auxquels sont attachés des dimensions conceptuelles diverses : le système schématique de la structure configurationnelle, celui de la perspective, celui de la distribution de l'attention, celui de la dynamique des forces, etc. Chacun de ces systèmes est organisé selon des principes. Le système de la perspective est celui qui présente un d'intérêt primordial pour l'analyse des verbes de perception sensorielle. Il traite de la perspective que l'on peut avoir d'une entité. Talmy (2000) définit ce système en termes visuels : « While this schematic system is presumably neutral to particular sensory modalities, it is most readily characterized in visual terms […].» (p. 68) Ce système comprend plusieurs catégories schématiques : la localisation de la perspective (qui décrit la localisation du point de perspective), la distance de perspective (la distance relative du point de perspective par rapport à l'entité), le mode de perception (le point de perspective est-il stationnaire ou en mouvement ?) et la direction (localisation du point de perspective par rapport à l'entité).

Il faut souligner que les catégories schématiques que sont le mode de perception, la direction… relèvent parfois de la combinaison de plusieurs facteurs issus de systèmes schématiques différents : par exemple, le facteur *focus*

attentionnel est issu du système de la distribution attentionnelle et il intervient dans la définition de la catégorie *direction de la perspective* relevant, elle, du système de la perspective.

Dans notre analyse, nous suivrons les principes émis par Talmy (2000), mais nous adapterons entre autres son système de la perspective à l'analyse des diverses modalités sensorielles. Il s'agira donc de définir les systèmes schématiques pour chaque mode de perception (la vue, l'ouïe, l'odorat, etc.) et d'établir les principes communs aux différentes modalités sensorielles. Nous procéderons à cette adaptation des systèmes schématiques en nous basant sur l'étude de la physiologie de chaque modalité sensorielle : la physiologie de la vue, de l'audition, du goût, etc.

III. 2.2. Physiologie de la perception sensorielle : le cas du verbe *entendre*

Il s'agit de relier la physiologie de la perception sensorielle, en d'autres termes, notre mode de cognition, à l'usage que la langue fait de l'expression de la perception sensorielle. Notre hypothèse est que notre mode de cognition est transposé dans le langage et donc dans notre façon de nous exprimer. Par exemple, l'étude de la physiologie de la perception auditive devrait guider les informations que le verbe *entendre* permet d'exprimer au moyen des arguments qu'il accepte. L'étude de la physiologie ne doit pas être vue comme une simple transposition de mécanismes biologiques et physiologique dans l'analyse linguistique, mais plutôt comme un ensemble de possibilités d'expression que l'on peut trouver dans le comportement du verbe étudié.

Nous présentons ici des bribes de l'analyse de la perception auditive en relation avec l'utilisation du verbe *entendre* pour montrer les possibilités qu'offre ce type d'approche. Il ne s'agit donc que d'une version synthétisée de l'analyse (Piron 2003). Dans ce document, nous nous centrerons d'ailleurs sur les cas de l'expression de la perception sensorielle et non sur les sens qui tombent en dehors de ce domaine.

Les données sur la base neuronale de l'audition (Carney 2002) et sur la perception et la cognition auditive (McAdams & Drake 2002) établissent des différences que nous abordons ci-dessous au travers de leur expression linguistique. Dans le cas de l'audition, la première donnée physiologique est la compétence dans ce domaine. Il est possible d'exprimer cette compétence : *il entend, il n'entend pas, il n'entend plus.* Au-delà de la constatation de la compétence, celle-ci peut être caractérisée par un certain degré d'acuité. La perte de l'acuité auditive monaurale ou biaurale que l'on définit par le terme d'hypoacousie (qui est souvent, dans les faits, de la presbyacousie) est courante. Linguistiquement, elle s'exprime au moyen d'une modification par un adverbe de degré : *il entend mal, il n'entend pas bien.* L'hypoacousie peut être due à un dysfonctionnement d'une partie de l'appareil auditif et s'exprime linguistiquement de manière monaurale puisque le traitement des signaux acoustiques est d'abord monaural. Ainsi, *je n'entendais presque plus rien de l'oreille droite.*

Des études récentes ont montré l'existence d'acuités auditives spécifiques. Les recherches ont décelé trois modes de perception auditive présentant des autonomies fonctionnelles : perception verbale (mode phonétique), perception environnementale (mode auditif) et perception musicale (mode musical). L'acuité musicale est exprimable en soulignant une harmonie : *il entend exactement, il entend juste*. De fait, la perception musicale repose sur la perception des harmoniques –des entiers multiples de la fréquence fondamentale (Justus & Bharucha 2002).

Au-delà de la compétence auditive, les études physiologiques relèvent la perception auditive. La perception auditive est en partie automatique, mais même dans ce cas l'apparition de l'attention se fait très tôt (Alain, Arnott & Picton 2001). Cette constatation permet de distinguer deux types d'attention auditive : l'attention spontanée et l'attention volontaire. L'attention spontanée est ce que l'on appelle le réflexe acoustique (McAdams S. & Drake 2002). Ce réflexe est présent pour les trois modes auditifs reconnus : mode auditif général (Mann & Liberman 1983), mode phonétique (Mann & Liberman 1983) et mode musical (Justus & Bharucha 2002).

Exemple 15. MODE AUDITIF. la propriétaire des chiens qui avait d'ailleurs été alertée en entendant **les tirs** (entendre_LaMeuse_201001_08)

Exemple 16. MODE PHONÉTIQUE. C'est en tout cas le sentiment qui ressort après avoir entendu **les premières réactions de Claude Willemart** que nous avons contacté (entendre_LaMeuse_271001_21).

Exemple 17. MODE MUSICAL. depuis l'ouverture de leur salle, nous ne savons plus dormir. Nous n'entendons pas **la musique mais les basses**. (entendre_LaMeuse_241001_09).

Si l'on s'attache maintenant aux différentes caractéristiques acoustiques que l'être humain est capable de percevoir, on se rend compte qu'on les retrouve exprimées de diverses manières autour du verbe *entendre.* Ainsi, à l'intérieur du syntagme nominal complément du verbe, on peut trouver un syntagme adjectival qui tantôt exprimera la séquencialisation du son entendu (exemple 18), la discrimination de l'ordre temporel des sons en séquence (exemple 19), l'intensité du son (exemple 20), le profil spectral (exemple 21), etc.

Exemple 18. En outre, des gardes frontières pakistanais ont dit avoir entendu **cinq explosions près d'un réseau de souterrains utilisé par l'organisation** (entendre_LaMeuse_261001_03)

Exemple 19 Alors qu'elle et ses collègues évacuaient la tour, ils ont entendu **une deuxième explosion**. (entendre_LaMeuse_101001_31)

Exemple 20. A deux cents mètres des lieux, les pompiers ont entendu **une énorme déflagration** (entendre_LaMeuse_221001_07)

Exemple 21. On y entend **deux parties contrastées** (entendre_LaMeuse_291001_04)

Un syntagme prépositionnel peut compléter le syntagme nominal et présenter la source du son, qu'il s'agisse d'un son environnemental (exemple 22), de paroles (exemple 23) ou encore de musique (exemple 24).

Exemple 22. Elle doit être bien lourde et surtout jamais fendue.{S} Je la secoue %% pour entendre **l'agréable clapotis de son lait**. (entendre_LaMeuse_181001_08)

Exemple 23. C'est en tout cas le sentiment qui ressort après avoir entendu **les premières réactions de Claude Willemart** que nous avons contacté à so (entendre_LaMeuse_271001_21)

Exemple 24. On aura l'occasion d'entendre **Steve Houben au sax soprano** (entendre_LaMeuse_031001_18**)**

De récentes études (Romanski et al. 1999) ont montré que, dans le cortex auditif, il existe des chemins séparés (mais interagissant) pour la description du *quoi* et du *où* des inputs auditifs. Les études menées tant sur les animaux que sur les humains suggèrent qu'il existe un schéma de traitement double : l'information concernant les patrons auditifs et les objets est traitée dans le gyrus temporel supérieur tandis que l'information auditive spatiale est traitée dans les régions pariétales du cortex. Dans l'expression linguistique de l'audition, on retrouve ce double aspect lorsqu'un syntagme nominal complexe contient, d'une part, le son entendu (tête du syntagme) et, d'autre part, un syntagme prépositionnel pour la localisation de ce son (exemples 25 et 26). On constate également que l'expression de la localisation du son tient compte de la position d'origine du son par rapport à l'auditeur.

Exemple 25. cteur décide alors d'attendre quelques minutes dans sa voiture. [Soudain, il entend **une voix à côté de sa portière** : "]C'est bien ici, docteur!". (entendre_LaMeuse_111001_14)

Exemple 26. sont heureusement sains et saufs! [M. Martin qui dormait au rez-de-chaussée a entendu **des crépitements en provenance de la cour**.] (entendre_LaMeuse_221001_08)

Bien d'autres remarques sont à faire sur le parallèle entre la perception physiologique et l'expression de cette perception. L'objectif que nous poursuivons dans notre étude est d'établir ces parallélismes pour obtenir une analyse ancrée dans la physiologie de la perception. Ces informations seront encodées dans le lexique que nous construirons pour les verbes de perception (section V. - lexique computationnel).

IV. FORMALISMES DE REPRÉSENTATION DE L'ANALYSE LINGUISTIQUE

IV. 1. PROBLÉMATIQUE ET OBJECTIFS

Une fois l'analyse linguistique menée, il faut la représenter dans un formalisme qui puisse décrire les phénomènes syntaxiques majeurs qui caractérisent le comportement des verbes que nous étudions. Il s'agit de décrire les chaînes de mots bien formées. Ce formalisme devrait également permettre d'identifier le sens que prennent les verbes de perception sensorielle dans des phrases spécifiques. Mais surtout il s'agit d'utiliser un formalisme qui, tout donnant une description précise de phénomènes relevant des langues naturelles, puisse en « donner une caractérisation qui soit interprétable par un ordinateur. » (Shieber 1990, p. 30)

La syntaxe est décrite dans une grammaire qui définit les principes et les contraintes qui régissent la combinatoire des mots. La première fonction de cette grammaire est de fixer les règles qui forment des phrases grammaticales. Sa seconde fonction est d'associer aux phrases des représentations syntaxiques qui explicitent leur structure. Dans le cas des analyseurs syntaxiques, l'utilisation de méthodes probabilistes est bien évidemment réalisable (Bod 1998 ?),

mais de telles analyses débouchent sur des grammaires peu réutilisables. Le problème du coût généré par l'annotation du corpus servant de base à ce type de grammaires est loin d'être négligeable. Dans le même domaine, les analyses linguistiques permettent d'obtenir une grammaire lisible. Le coût relié à ce type d'approches est, cette fois, relié à la construction de la grammaire. L'avantage est qu'elle est réutilisable, lisible et réversible (utilisable en analyse et en génération).

Le développement d'un lexique sert à encoder le sens et à guider l'analyse syntaxique. C'est dans le lexique que peuvent se trouver selon les formalismes les informations sur la morphologie, la sémantique, une partie de la syntaxe et aussi la phonétique.

Les formalismes qui peuvent servir à la représentation linguistique sont les grammaires formelles. Chomsky (1957, 1959, 1965) définit ces grammaires comme des quadruplets comprenant des catégories syntaxiques (nœuds non terminaux), des mots (nœuds terminaux), des règles de production (par exemple, SV $\rightarrow$ V SN) et un symbole de départ (par exemple, P). Il existe plusieurs types de grammaires formelles[18] qui n'ont pas toutes le même pouvoir génératif. Cela dépend des règles de production qu'elles autorisent. Les plus utilisées sont les grammaires de type 2. En effet, elles sont assez puissantes pour décrire la plupart des structures syntaxiques de la langue. Elles sont cependant limitées pour décrire certains phénomènes (l'accord, surtout dans les cas de phénomènes non locaux comme *elle persuade ses fils de devenir danseurs*) [Bouillon et al. 1998]. Elles permettent de reconnaître les phrases grammaticales et de leur assigner une représentation (arbre). Par ailleurs, elles ont une interprétation déclarative (indépendante de toute notion d'ordre).

Dans le courant des années 80, un mouvement est apparu en réaction au courant transformationnel de Chomsky et surtout pour pallier les limites formelles de l'implémentation de ce formalisme. Une constellation de formalismes plutôt qu'une théorie ont été développés : la grammaire catégorielle (Moortgat 1988; Steedman 1996, 2000), la grammaire d'unification fonctionnelle (FUG – Kay 1984), la grammaire lexicale fonctionnelle (LFG – Bresnan 1982, 2001), la grammaire syntagmatique généralisée (GPSG – Gazdar & Pullum 1985), la grammaire syntagmatique guidée par les têtes (HPSG – Pollard & Sag 1994, Sag & Wasow 1999) et la grammaire d'arbres adjoints (TAG – Joshi & Schabes 1992). Ces formalismes sont le fruit de recherches distinctes en linguistique informatique, en linguistique formelle et en traitement automatique du langage. Tous les formalismes cités utilisent l'unification, un outil puissant. Mais chacun le fait à sa manière, notamment en ajoutant des mécanismes spécifiques. C'est essentiellement l'utilisation de l'unification qui regroupe ces divers formalismes grammaticaux (Shieber 1986). Les formalismes grammaticaux les plus utilisés actuellement pour le TAL appartiennent à cette famille des grammaires d'unification (aussi appelées

[18] Les grammaires de type 0 n'imposent aucune restriction sur les règles de production. Les grammaires de type 1 (dépendantes du contexte) imposent une restriction sur la taille des règles (le côté droit ne peut pas avoir moins de symboles que le côté gauche). Les grammaires de type 2 (indépendantes du contexte) ont un symbole non terminal à gauche et une chaîne non nulle à droite. Enfin, les grammaires de type 3 (régulières) ont soit un symbole non terminal à gauche et à droite soit un symbole terminal unique soit un symbole terminal suivi d'un non terminal (pour une grammaire régulière à droite). (Gurari 1989)

grammaires à base de contraintes). En font usage les projets LinGo, XTAG, LTAG et Graal. À part l'unification, les autres caractéristiques communes à ces formalismes sont le rejet des transformations, la réhabilitation des descriptions de surface, la formalisation à base de structure de traits, l'appartenance aux formalismes déclaratifs et la séparation stricte entre données linguistiques et programme de traitement (permettant leur utilisation en analyse, en génération et en désambiguïsation).

Dans notre étude, nous avons opté pour la grammaire d'unification HPSG (*Head-driven Phrase Structure Grammar*). Les possibilités qu'elle offre pour la description du lexique en font un formalisme très qualifié pour l'étude que nous menons. Par ailleurs, des environnements de programmation permettent son implémentation. Dans notre cas, nous utiliserons l'environnement LKB (*Linguistic Knowledge Building*). Notre objectif consiste donc à transposer la description linguistique des verbes de perception sensorielle en français dans le formalisme HPSG tel qu'il peut être implémenté dans LKB.

IV. 2. LE MODÈLE HPSG

HPSG est une théorie monostratale non dérivationnelle (Blache 2001) : les relations entre éléments dans une structure linguistique ne sont pas analysés à l'aide de mouvements ou de transformations (cas de la théorie du liage), mais en termes de partage de l'information. Il n'y a qu'un seul niveau de représentation. C'est une théorie qui a surtout la particularité, par rapport aux autres formalismes d'unification, d'intégrer des connaissances à la fois syntaxiques et sémantiques. Le mécanisme fondamental de la description des relations syntaxiques entre constituants est celui de la propagation de traits. Ceci est possible grâce au cadre formel dans lequel le formalisme exprime les informations : les structures de traits.

IV. 2.1. Les catégories

HPSG s'écarte des approches qui dissocient les informations linguistiques dans des modules séparés (modèle chomskyen ou LFG). Dans HPSG, on retrouve l'approche saussurienne selon laquelle un signe linguistique est une combinaison de caractéristiques phonétiques, syntaxiques et sémantiques. L'encodage des informations linguistiques dans HPSG en particulier et dans tous les formalismes d'unification en général se fait par des structures de traits[19] (désormais, ST). Le cadre formel duquel est issue la ST est celui de la logique des attributs-valeurs (Pereira 1985; Shieber 1984, 1986; Johnson 1988). Une ST est un ensemble de paires d'attributs - valeurs. Ainsi [catégorie SN] est une ST qui associe un attribut (*catégorie*) à une valeur (*SN*). Chaque attribut est un symbole atomique issu d'un ensemble fini d'attributs. Quant aux valeurs, ce sont soit un symbole atomique, soit une autre structure de traits. L'illustration se fait souvent sous la forme d'une matrice d'attributs-valeurs (AVM – *Attribute Value Matrix*) :

[19] L'utilisation des traits en linguistique remonte vraisemblablement à Jakobson (1939) [Jurafsky & Martin 2000]. L'utilisation des traits distinctifs se pratique aussi beaucoup en phonologie. Leur utilisation s'est faite en sémantique et aussi en syntaxe (Chomsky 1965 notamment). Mais leur utilisation en tant que structures de traits relève des formalismes d'unification.

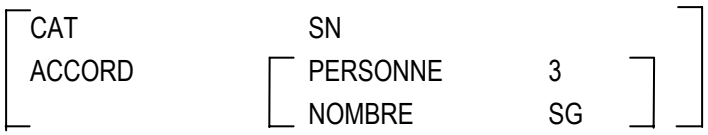

Figure 3. Une structure de traits.

Ces ST peuvent être complexes : la liste des attributs et de leur valeur peut être longue, mais elle peut aussi présenter des valeurs partagées. Ce mécanisme de partage des valeurs est central en HPSG. Il permet de signifier que des attributs partagent les mêmes valeurs. Ceci est très utile et puissant pour représenter des phénomènes d'accord ou de dépendance de longue distance notamment. Une ST sous forme matricielle est l'équivalent formel d'un chemin de traits, c'est-à-dire une liste de traits qui mène à une valeur. L'équivalent graphique de cette représentation de la ST est le graphe acyclique orienté (DAG – *Directed Acyclic Graph*). Dans ce cas, les traits sont des arcs étiquetés et les valeurs sont des nœuds.

Outre le fait qu'une ST est implémentable, elle présente d'autres avantages très importants (Pereira 1985) : elle regroupe sur le même plan des informations de différentes natures (notamment syntaxiques et sémantiques en HPSG), elle affine les contraintes du lexique et permet par conséquent de simplifier les règles de grammaire et, enfin, elle enrichit les représentations linguistiques. Les catégories utilisées en HPSG en particulier ne sont pas de simples étiquettes *SN, SP,* etc. Si on préserve l'ensemble des catégories grammaticales (SN, SP, etc.), on établit en même temps des distinctions entre les membres d'une même catégorie puisqu'on en fait des représentations complexes de généralisations syntaxiques et sémantiques. Elles comprennent des notions comme la sous-catégorisation, les traits d'accord, etc. Finalement, l'équivalence des ST avec les DAG permet beaucoup plus de souplesse et de puissance dans l'expression du formalisme puisqu'il n'y aucune contrainte sur l'ordre des arcs du graphes et la réentrance permet de lier des sous-structures très facilement.

IV. 2.2. Les règles linguistiques

Le formalisme exprime tous les objets linguistiques par des ST. C'est ainsi que les règles linguistiques sont elles aussi vues comme des objets qui ont des propriétés. L'information sur ces propriétés est représentée par des contraintes. HPSG possède des schémas de règles qui définissent les contraintes de bonne formation pour les relations hiérarchiques. Si l'on prend cela dans une perspective de représentation arborescente, ces schémas donnent les règles de bonne formation des arbres locaux. Il y a ainsi 6 schémas qui exposent les types de relations entre une racine et ses nœuds filles (Sag & Wasow 1999). Par exemple, un schéma indique que la racine est un signe sous-catégorisant un sujet (ou un spécifieur) et que la fille tête est un signe lexical. En plus des schémas, il existe des principes qui encodent les contraintes que certains types de structures doivent satisfaire. Par exemple, le principe de valence rend compte de la sous-catégorisation. Le principe indique comment retirer des constituants

réalisés à partir d'une liste contenue dans le trait VAL (valence)[20]. De tels principes vérifient qu'une structure décrite a bien la structure donnée et complètent les structures qui ne sont que partiellement connues.

IV. 2.3. Le lexique

HPSG est une théorie fortement lexicalisée. Le lexique prend donc en charge la représentation d'un grand nombre d'informations à la fois syntaxiques et sémantiques. Il n'y a pas de modules séparés pour la syntaxe et la sémantique, mais bien intégration des divers types de données. Comme pour le reste des objets linguistiques, HPSG représente les entrées lexicales par des ST. HPSG s'intègre dans le mouvement d'évolution des théories linguistiques vers une plus grande lexicalisation. Davantage d'informations sont désormais contenues dans le lexique. Cette importance prise par le lexique conduit à une autre forme de modularisation. Il ne s'agit plus de distinguer des modules syntaxiques et sémantiques, mais d'avoir un module contenant toutes les informations sur les mots et ensuite des principes gérant la cooccurrence de ces mots (d'une part, les principes comme celui de la valence et, d'autre part, les schémas). Lorsqu'il s'agit d'implémenter HPSG dans l'environnement LKB, on se trouve devant le problème de la constitution d'un lexique computationnel très riche. Cette richesse du lexique est nécessaire en TAL pour pouvoir fournir des informations précises et utile à l'analyse, mais ce lexique doit être structuré et implémentable efficacement (Briscoe 1992).

Les approches computationnelles du lexique l'ont longtemps considéré comme une liste de mots sans liens entre eux, comme autant d'entrées qu'il y a d'ambiguïtés. Ainsi, si un verbe a deux emplois (transitif et intransitif), il aura alors deux entrées. Cette approche ne rend pas compte du fait qu'un grand nombre d'ambiguïtés sont régulières : contenant / contenu, viande / animal, etc. Cela va à l'encontre de l'objectif qui vise à rendre le lexique compact. Et cela peut d'ailleurs générer des ambiguïtés. Par ailleurs, une vision psycholinguistique du lexique réfute les longues énumérations.

Les théories lexicales linguistiques tentent de représenter le lexique en exprimant les liens qui unissent les différents sens d'un mot. On distingue essentiellement trois approches. L'approche relationnelle insiste sur les relations qui existent entre les concepts lexicaux (Fellbaum 1998) : organisation des mots du lexique dans des ensembles de synonymes, avec des concepts lexicalisés sous-jacents. Dans l'approche compositionnelle, les traits sémantiques se combinent. Jackendoff (1990) décrit un ensemble fini de primitives mentales et un ensemble fini de principes de combinaisons mentales qui, ensemble, permettent de former le lexique par composition, par combinaison. Il est possible d'adopter une approche mixte, qui combine relationnalité et compositionnalité. C'est l'approche que propose Pustejovsky (1995). C'est cette conception du lexique qui sert de base à l'implémentation du lexique dans le système LKB. Pustejovsky adopte une position plus souple que Jackendoff envers les primitives. Pour lui, les aspects génératifs sont plus importants que la décomposition en un nombre fini de primitives. Sa description concerne à la

[20] Ce trait (Sag & Wasow 1999) remplace le trait SUBCAT des versions précédentes de HPSG (Pollard & Sag 1994).

fois le niveau lexical des mots, mais aussi leur composition et leur variation en contexte. Il s'agit d'une représentation du lexique à plusieurs couches qui incorporent les aspects les plus saillants de la connaissance du monde relative à un mot : la structure argumentale, la structure événementielle (Vendler 1967, Dowty 1979) et la structure qualia (les attributs qui définissent un objet telle son utilité, ses parties…). Pustejovsky prône une représentation plus fine que celle des rôles thématiques (Fillmore 1968, 1977). Il développe également les opérations qui permettent de relier les différents niveaux de description du lexique qu'il postule.
Dans le système LKB, le lexique –comme n'importe quel objet– est exprimé par des listes d'attributs-valeurs qui s'insèrent ainsi parfaitement dans un formalisme d'unification et peut donc utiliser l'unification, c'est-à-dire l'opération générale du formalisme (section IV. 2.4.).

L'expression des régularités au sein du lexique peut se faire en construisant un lexique dit d'héritage (Copestake 1992, Pustejovsky 1991, 1995). Dans ce cas, on considère le lexique comme une hiérarchie de traits. Différents types d'héritage sont possibles : l'héritage simple accepte l'héritage d'un seul nœud à la fois; l'héritage multiple rend possible l'héritage de plusieurs nœuds à la fois. Dans ce cas, la hiérarchie est un treillis. Cette notion est attrayante d'un point de vue de traitement automatique, car les opérations posées sur le lexique peuvent alors être formalisées en termes de notions théoriques relatives aux treillis (Carpenter 1992). Et d'un point de vue linguistique, le treillis permet d'exprimer la nature hiérarchique du lexique. Il est ainsi possible de représenter les ST du lexique selon des relations de subsomption (au plus un nœud est haut dans la hiérarchie, au plus il est général). De cette manière, l'information qui se répète au travers de la hiérarchie n'est spécifiée qu'une seule fois (Flickinger 1987, Pollard & Sag 1987, Sanfilippo 1993). Peuvent ainsi être spécifiées comme patrons des propriétés communes à tous les verbes (présence d'un sujet, partie du discours, etc.) ou à des classes de verbes (les verbes transitifs, les verbes de perception sensorielle, etc.). L'héritage par défaut permet d'atteindre un niveau de compression du lexique qui permet d'exprimer des généralisations (Flickinger 1987, Gazdar 87). Malheureusement, des problèmes se posent rapidement lorsqu'il y a des héritages par défaut multiples. Pour gérer ces problèmes d'héritage, il est possible d'expliciter l'ordre de priorité ou d'appliquer des solutions *ad hoc*. Aucune solution parfaitement viable et générale n'a pu être proposée.
Le traitement de l'ambiguïté lexicale est un autre défi pour l'organisation des lexiques computationnels. La création de représentations lexicales à plusieurs couches (Pustejovsky & Boguraev 1993, Pustejovsky 1995) rend possible de regrouper les différents usages d'un même mot dans une méta-entrée que les règles lexicales peuvent utiliser (Copestake & Briscoe 1995). Une autre possibilité que nous étudierons avec intérêt est celle de Sanfilippo & Benkeremini (1994) et Sanfilippo (1995). La proposition consiste à exprimer un mot polysémique comme un objet lexical polymorphe, c'est-à-dire une ST contenant plusieurs extensions qui décrivent les usages possibles du mot. Les ambiguïtés lexicales du mot polysémique peuvent être levées de manière déterministe à l'aide de l'information contextuelle à la fois syntaxique et sémantique.

Tel qu'exprimé jusqu'à présent, le lexique présente encore des problèmes pratiques. Il pourrait permettre une ST

inadéquate telle [nombre féminin]. En effet, il n'y a aucune contrainte sur les attributs et leurs valeurs possibles. Une solution à ce problème est l'utilisation de types. Les ST typées ont été proposées indépendamment par Calder (1987) et par Pollard & Sag (1987) et elles ont été formalisées par Carpenter (1992). Les caractéristiques du système des types sont au nombre de trois. D'abord, chaque ST est étiquetée par un type. En HPSG, toute ST devra donc indiquer de quel type elle relève : *sign, lexical-sign* et *phrasal-sign* (Sag & Wasow 1999). Ensuite, chaque type présente des conditions qui expriment les traits qui lui sont appropriés. Par exemple, la ST typée pour le *verbe* devra contenir le trait *accord,* le trait *valence,* le trait *vform*; ce que ne contiendra pas la ST typée pour le nom. Celle-là aura seulement le trait *accord*. Il y a donc une nouvelle classe d'objets dans le lexique de HPSG : les ST typées. Enfin, les types sont organisés en hiérarchie et les types plus spécifiques héritent des propriétés des types plus abstraits (Lascarides & Copestake 1997).

Un aspect en faveur de l'utilisation des ST typées est la possibilité de donner ainsi des informations sur un objet linguistique sans pour autant lui affecter des valeurs spécifiques. C'est ce que l'on appelle aussi une contrainte. La notion de contrainte joue un rôle important en informatique. Les langages de programmation par contraintes (Jaffar 1986, Colmerauer 1987, Cohen 1990) reposent sur le mécanisme général de la satisfaction de contraintes. La contrainte permet de spécifier le domaine de définition d'une variable, en informatique comme en linguistique[21] d'ailleurs. Les catégories linguistiques et les règles linguistiques sont alors vues comme des objets possédant des propriétés. L'information sur ces propriétés est représentée par des contraintes. Le modèle HPSG appartient ainsi aux formalismes à base de contraintes. Une information à granularité plus fine telle que l'envisage HPSG s'avère nécessaire pour traiter des phénomènes linguistiques complexes. La compilation efficace de la grammaire et de la sémantique bloque les problèmes de sur-génération et d'acceptation trop large. De plus, les rôles sémantiques des arguments sont apparaissent dans la ST[22]. Ainsi, *entendre une orange* peut être bloqué par un formalisme de contraintes.

IV. 2.4. L'opération d'unification

La technique informatique qui permet de travailler sur les ST est l'unification. Il s'agit d'une opération fondamentale et puissante qui permet de rejeter des ST incompatibles et d'accepter celles qui sont compatibles. L'unification est un opérateur binaire ($\cup$) qui accepte deux ST en arguments et retourne une ST unifiée si l'opération a réussi.

[NOMBRE SG] $\cup$ [NOMBRE PL] = échec

[NOMBRE SG] $\cup$ [NOMBRE [ ]] = [NOMBRE SG]

---

[21] La notion de contrainte est très répandue en linguistique, mais toutes les théories qui en font usage ne reposent pas effectivement sur les contraintes. La théorie qui fait le plus usage des contraintes est la théorie de l'optimalité. Les grammaires génératives évoluent dans cette direction : la théorie des principes et paramètres, de même que le programme minimaliste (Chomsky 1995) reposent sur la notion de contrainte en proposant un certain nombre de contraintes générales.

[22] Ces rôles argumentaux sont cependant réduits à leur plus simple expression. Ils relèvent en fait de la sémantique des situations (Barwise & Perry 1983). Cet aspect sémantique demande à être développé.

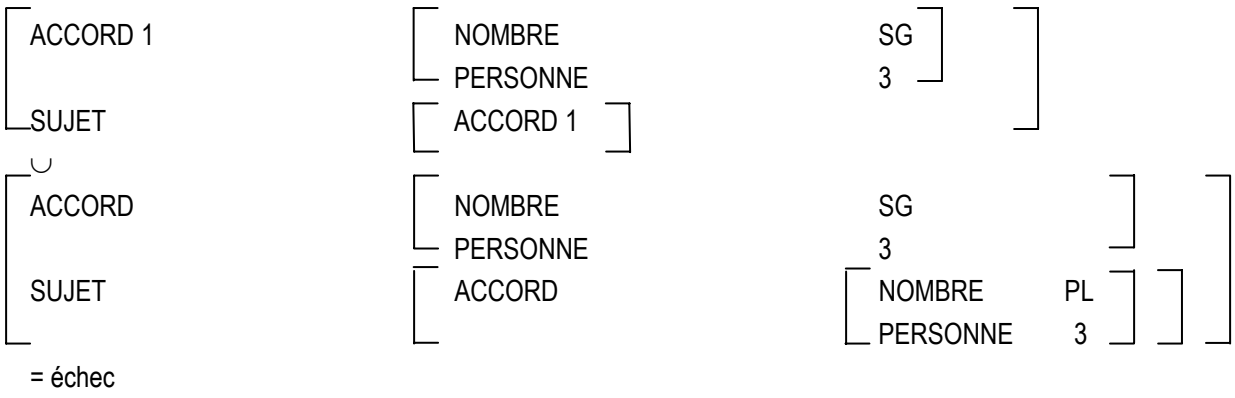

Figure 4. L'opération de l'unification de deux structures de traits.

'unification de deux ST A et B (A ∪ B) retourne la structure minimale qui est à la fois une extension de A et de B. En d'autres termes, l'unification ajoute de l'information. Si une telle structure unifiée est impossible, l'unification échoue et la chaîne de mots proposée (correspondant aux ST à unifier) est rejetée. Ainsi, on peut rejeter **il entendent*, mais il est aussi possible de rejeter des phrases sémantiquement inacceptables si le lexique contient les traits et les valeurs permettant de le faire : *ils entendent une orange.*

V. TRAITEMENT AUTOMATIQUE DE LA REPRÉSENTATION LINGUISTIQUE

Le formalisme HPSG, un formalisme à base de contraintes, permet un traitement automatique. En effet, l'expression de ses objets, qu'ils soient des éléments du lexique, des principes grammaticaux ou sémantiques, sont tous exprimés sous la forme de ST. Cette formalisation est encodable. Et surtout, l'opération pouvant manipuler ces objets est une opération informatique en usage dans d'autres domaines que celui de la linguistique informatique.

Dans notre étude, l'utilisation du formalisme HPSG pour le traitement automatique se fera grâce au système LKB (*Linguistic Knowledge Building*). Il s'agit d'un environnement de développement de grammaires et de lexiques pouvant être utilisés avec n'importe quel formalisme à base de contraintes. Il est plus spécifiquement adapté aux ST typées. Plus techniquement, le système LKB est implémenté en Common Lisp.

Le système LKB permet de faire à la fois du parsage et de la génération de phrases. Jusqu'à présent il a essentiellement été testé avec des grammaires basées sur HPSG, mais il est censé être indépendant de tout cadre théorique précis.

Les grammaires et les lexiques disponibles actuellement relèvent du projet LinGO pour l'anglais. La contribution de ce travail est donc de construire de petits modules pouvant fonctionner sur des phrases en français.

VI. DIMENSIONS COGNITIVE ET INFORMATIQUE

La dimension cognitive du projet intervient à différents niveaux. Tout d'abord, l'objet d'étude relève de la linguistique puisqu'il s'agit des verbes de perception sensorielle en français. Leur étude se fait dans le cadre d'une analyse empirique (analyse basée sur un corpus) et rationnelle (basée sur l'intuition), en analysant les sens qui sont reliés aux différentes constructions syntaxiques dans lesquelles les verbes peuvent apparaître. L'étude des préférences sélectionnelles des verbes et de leurs patrons syntaxiques est complétée par une analyse sémantique qui s'insère dans le cadre de la sémantique cognitive. La particularité de notre étude est de fonder l'analyse sémantique sur la physiologie de la perception sensorielle (perception auditive, gustative, etc.). L'analyse porte sur un phénomène langagier (les verbes de perception sensorielle en français) qui a la particularité d'être l'expression de plusieurs de nos modes de cognition du monde qui nous entoure. En plus, cette analyse repose sur la physiologie de cette cognition puisqu'elle utilise des données issues de la neurologie et de la cognition de la perception pour expliquer l'organisation de cette perception dans le système du langage.

La dimension informatique de ce projet intervient, elle aussi, à différents niveaux. L'analyse linguistique basée sur corpus requiert la construction de ressources consultables électroniquement. Leur encodage en XML présente une réflexion sur l'organisation d'un tel document devant servir de base de données. Enfin, la représentation des connaissances issues de l'analyse linguistique se fait dans un formalisme à base de contraintes : HPSG. Ce formalisme est implémenté dans le système d'environnement LKB, qui permet l'analyse et la génération de phrases. Une tâche importante consiste en la construction du lexique computationnel que cette application nécessitera. La représentation de ce lexique sera l'aboutissement de l'analyse linguistique, mais aussi l'intégration de réflexions sur le typage et l'héritage des structures de traits.

VII. CALENDRIER

VII. 1. Description linguistique

Champs sémantiques :	Ils sont établis de manière générale. La compilation des notices de dictionnaires anciens et modernes est terminée, leur analyse a d'ailleurs déjà fait l'objet de communications (Piron 2002a, 2002b, 2003).
Récupération des données pour la constitution du corpus :	Étape terminée.
Choix des sigles d'annotation, des niveaux d'analyse, du schéma XML :	Étape terminée.
Encodage XML du corpus, avec annotations :	Le verbe *entendre* est terminé.

	Les trois autres verbes (*voir, sentir, goûter*) : 01/2004 à 02/2004
Gestion du document XML	03/2003

VII. 2. Analyse linguistique

Restrictions sélectionnelles et sens des verbes :	*Entendre* est terminé (Piron 2004, à paraître). Les autres verbes : 03/2004 à 06/2004

VII. 3. Formalisme de représentation linguistique

Analyse du modèle HPSG pour l'encodage des verbes de perception sensorielle :	07/2004 à 08/2004.

VII. 4. Traitement automatique de la représentation linguistique

Formalisation du lexique computationnel et des schémas et principes pour l'analyse des verbes de perception sensorielle :	09/2004 à 12/2004.
Dépôt de la thèse :	12/2004

VIII. BIBLIOGRAPHIE

ABEILLÉ A. & BLACHE P. (2000). Grammaires et analyseurs syntaxiques. In Pierrel J.-M. *Ingénierie des langues,* Hermès, pp. 51-76.

ABEILLÉ A. & CLÉMENT L. (1999). A tagged reference corpus for French, *LINC'99 Proceedings*, EACL workshop, Bergen.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003) Building a Treebank for French. In ABEILLÉ A. (ed.) *Building and Using Large Parsed Corpora*. Language and Speech series, Kluwer, pp. 165-187.

ABEILLÉ A., CLÉMENT L., KINYON A. &TOUSSENEL F. (2001). The Paris 7 annotated corpus for **French** : some experimental results, in WILSON (ed) *Corpus Linguistics*, Lancaster

AKMAJIAN A. (1977) The Complement structure of Perception Verbs in an Autonomous Framework. In P. Culicover *et al*. *Formal Syntax,* pp. 427-460.

ALAIN C.A., ARNOTT S. R. & PICTON T.W. (2001) Bottum-up and Top-down influences on auditory Scene Analysis : Evidence from Event-Related Brain Potentials. In *Journal of Experimental Psychology : Human Perception and Performance,* 27(5), pp. 1072-1089.

ALLEN J. (1987) Natural Language Understanding. Benjamin Cummings Publishing Co.

ATKINS T. S. & LEVIN B. (1992) Admitting impediments. In *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*. Lawrence Earlbaum, Hillsdale.

BARWISE J. & PERRY J. (1983) *Situations and Attitudes*. Cambridge, MIT Press.

BLACHE P. (2001) *Les grammaires de propriétés. Des contraintes pour le traitemetn automatique des langues naturelles.* Hermès.

BOD R. (1998) Beyond Grammar: An Experience-Based Theory of Language. CSLI Publications, Cambridge University Press.
BONHOMME P. (2000) Codage et normalisation de resources textuelles. In PIERREL J. –M. *Ingénierie des langues,* Hermès, pp. 173-192.
BOUILLON P. *et al.* (1998) *Traitement automatique des langues naturelles.* Duculot.
BRANTS S. & HANSEN S. (2002) Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of theThird Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, pp.1643–1649.
BREKKE M. (1988) The Experiencer constraint. *Linguistic Inquiry,* 19, pp. 169-180.
BRESNAN J. (2001) *Lexical-functional syntax*. Blackwell.
BRESNAN J. (ed) (1982) *The mental representation of grammatical relations.* MIT Press.
BRISCOE E. J. Lexical issues in natural language processing. In E. Klein and F. Veltman, editors, *Natural Language and Speech*, pages 39--68. Springer-Verlag, 1992.
CALDER J. (1987) Typed Unification for Natural Language Processing. In Kahn G., MacQueen D. & Plotkin G. (eds) *Categories, Polymorphism, and Unification.* Centre for Cognitive Science, University of Edinburgh.
CAPLAN D. (1973) A note on the abstract readings of verbs of perception. *Cognition 2,* pp. 269-277.
CARNEY L. H. (2002) Neural Basis of Audition. In PASHLER H. *Steven's Handbooh of Experimental Psychology. Volume 1: Sensation and Perception*, John Wiley & Sons, pp. 341-396.
CARPENTER B. (1992) *The Logic of Typed Featured Structures*. Cambridge University Press.
CHAMBREUIL M. (sous la dir.) (1998). *Sémantiques*. Paris, Hermès.
CHOMSKY N. (1956) Three models for the description of language. *IRE Transactions on Information Theory,* IT-2(3), pp. 113-124.
CHOMSKY N. (1957) *Syntactic Structures*. The Hague, Mouton.
CHOMSKY N. (1959) On certain formal properties of grammars. In *Information and Control*, 2(2), pp. 137-167.
CHOMSKY N. (1965) *Aspects of the theory of syntax.* MIT Press.
CHOMSKY N. (1995) *The minimalist program.* MIT Press.
CHURCH K. & MERCER R. (1993) Introduction to the special issue on computation linguistics using large corpora, *Computational Linguistics,* 19 (1), pp. 1-24.
CLARK S. & WEIR D. (2002) Class-Based Probability Estimation using a Semantic Hierarchy, *Computational Linguistics*, 28(2), pp.187-206.
COHEN J. (1990)Constraint Logic Programming Languages, *Communications of the ACM, 33:7.*
COLMERAUER A. (1987) Opening the Prolog III Universe, *Byte*, August 1987.
COOPER W.E. (1974) Syntactic flexibility among English sensation referents. *Linguistics* 133, pp. 33-38.
COOPER W.E. (1975) Primacy Relations among English sensation referents. *Linguistics* 137, pp. 5-12
COPESTAKE A. & BRISCOE T. (1995) Semi-productive polysemy and sense extension, *Journal of Semantics*, 12, pp.15-67.
COPESTAKE A. & LASCARIDES A. (1997) Integrating symbolic and statistical representations: the lexicon-pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (ACL-EACL 97), Madrid, pp. 136-143.
COPESTAKE A. (1992) The ACQUILEX LKB: Representation Issues in the Semi-automatic Acquisition of Large Lexicons. In Antonio SANFILIPPO A. (ed) *The (other) Cambridge ACQUILEX papers*, University of Cambridge Computer Laboratory, Technical report No. 253.
COPESTAKE A. (2002) Implementing Typed Feature Structure Grammars. CSLI.
DARMESTETER A. (1887) *La vie des mots, étudiée dans leurs significations*. Paris.
DELSARTE P. & THAYSE A. (2001) *Logique pour le traitement de la langue naturelle.* Hermès, Science.
Dictionnaire de l'Académie française. (1694). Paris.
Dictionnaire de l'Académie française. (1798). Paris.
Dictionnaire de l'Académie française. (1835). Paris.

Dictionnaire de l'Académie française. (1877). Paris.
DOWTY D. (1979) *Word Meaning in Montague Grammar*. Reidel, Dordrecht.
EAGLES [Expert Advisory Group On Language Engineering Standards] (1996a) *Morphosyntactic Annotation.* EAGLES document EAG-CSG/IR-T3.1. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica ComputazionaleEAG---TCWG---SASG/1.8
EAGLES [Expert Advisory Group On Language Engineering Standards] (1996b) *Syntactic Annotation: Survey of Annotation practices*. EAG EAG-TCWG-SASG/2. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale
ESTIVAL D. & NICHOLAS N. (1999) TEI Encoding and Syntactic Tagging of an old French Text. *Computers and Humanities,* 33, pp. 155-174.
FAUCONNIER G. (1985). *Mental Spaces.* Cambridge, MA: MIT Press.
FELLBAUM C. 1998. *An Electronic lexical Database*. Cambridge, MIT Press
FELSER C. (1999) *Verbal Complement Clauses : a Minimalist Study of Direct Perception Constructions*. Benjamins.
FENSEL D., HENDLER J., LIEBERMAN H. & WAHLSTER W. eds. (2003) *Spinning the Semantic Web. Bringing the World Wide Web to its Full Potential.* Cambridge – London, MIT Press.
FILLMORE C. J. (1968) The case for case. In BACH E. & HARMS R. T. (eds) *Universals in linguistic theory*. Holt, Rinehart & Winston Inc.
FILLMORE C. J. (1977) The case for case reopened. In COLE P. & SADOCK J. M. (eds) *Syntax and Semantics 8: Grammatical Relations*. Academic Press, New York, pp. 59--81.
Flickinger D. (1987) *Lexical Rules in the Hierarchical Lexicon*. PhD thesis, Stanford University.
GARSIDE R., LEECH G.N & MCENERY T. (1997) *Corpus annotation. Linguistic information from computer text corpora.* London, Longman.
GAZDAR G (1987) Linguistic applications of default inheritance mechanisms. In WHITELOCK P. H., SOMERS H., BENNET P., JOHNSON R. &MCGEE WOOD M. (ed) *Linguistic Theory and Computer Applications*, pp. 37-68. Academic Press.
GAZDAR G**,** KLEIN E., PULLUM G. & SAG I. A. (1985) *Generalized Phrase Structure Grammar*. Blackwell.
GODEFROY F. (1880-1902) *Dictionnaire de l'ancienne langue française et de tous ses dialectes, du IXe au XVe siècle : composé d'après le dépouillement de tous les plus importants documents manuscrits ou imprimés, qui se trouvent dans les grandes bibliothèques de la France et de l'Europe et dans les principales archives départementales, municipales, hospitalières ou privées***.** Paris, 10 vol.
GURARI E.M. (1989) *An introduction to the theory of computation*. Computer Science Press.
HABERT B., NAZARENKO A. & SALEM A. (1997) *Les linguistiques de corpus.* Colin.
HAJIC J. (1998) Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*, Karolinum, pp. 106–132.
HUGUET E. (1966) *Dictionnaire de la langue française du seizième siècle.* Paris, Didier, 7 vol.
Jackendoff R. (1990) *Semantic structures.* MIT Press.
JAFFAR J. & LASSEZ J. – L. (1986) Constraint Logic Programming, *Research Report,* Department of computer Science, Monash University.
JAKOBSON R. (1939) Observations sur le classement phonologique des consonnes. In Blancquaert E., Pée W. (eds) *Proceedings of the Third International Congress of Phonetic Sciences,* Gent, pp. 34-41.
JOHNSON M. (1988) *Attribute-Value Logic and the Theory of Grammar.* CSLI.
JOSHI A. K. & SCHABES Y. (1992) Tree-adjoining grammars and lexicalized grammers. In *Tree Automata and LGS*. Elsevier Science, Amsterdam.
JURAFSKY D. & MARTIN J. H. (2000) *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.
Justus T. C. & Bharucha J. J. (2002) Music Perception and Cognition. In In PASHLER H. *Steven's Handbooh of Experimental Psychology. Volume 1: Sensation and Perception*, John Wiley & Sons, pp. 453-492.
KAY M. (1984) Functional unification grammar: a formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford University, California, July

1984. ACL.
KIRSNER R.S. (1979) On the passive of sensory verb complement sentences. *Linguistic Inquiry,* 8, pp. 173-179.
LAKOFF G (1990). *Talking Power. The politics of Language in our Lives*. New York.
LAKOFF G. & M. Johnson (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.
LAKOFF G. (1987) *Women, Fire and Dangerous Things.* Chicago: University of Chicago Press.
LANGACKER R. W. (1987). *Foundations of Cognitive Grammar,* vol. 1: *Theoretical Prerequisites.* Stanford: Stanford University Press.
LANGACKER R. W. (1991). *Foundations of Cognitive Grammar,* vol. 2: *Descriptive Application.* Stanford: Stanford University Press.
LANGACKER R. W. (2000). *Grammar and Conceptualization.* Mouton de Gruyter.
LANGÉ J. –M., LEMAÎTRE F. & PAPIN A. (1994). *Corpus bilingue EN-FR annoté. Notes sur les classes grammaticales et syntaxiques*. IBM Paris.
LAROUSSE P. (1903) *Dictionnaire complet et illustré de la langue française.* Paris.
LEBART L. & SALEM A. (1994) *Statistique textuelle*. Dunod.
LEEK F. van der & JONG A. (1982) The complement structure of perception verbs in English. In Daalder S. & Gerritsen (eds) *Linguistics in the Netherlands 1982,* Amsterdam, pp. 103-114
LEVIN B. (1993) *English Verb Classes and Alternations. A Preliminary investigation.* University of Chicago Press.
LITTRÉ É. (1872 et 1877) *Dictionnaire de la langue française.* Paris, 7 vol.
LYONS J. (1977) Semantics, vol. 1 & 2.Cambridge, CUP.
MANN V. & LIBERMAN P. (1983) Some differences between phonetic and auditory modes of perception. In *Cognition, 14,* 211-235.
MANNING C. D. & SCHÜTZE H. (2000) *Foundations of Statistical Natural Language Processing*. MIT Press.
MARTIN R. (1992) *Pour une logique du sens*, Paris, PUF
McAdams S. & Drake C. (2002) Auditory Perception and Cognition In PASHLER H. *Steven's Handbooh of Experimental Psychology. Volume 1: Sensation and Perception, John Wiley & Sons, pp. 397-452.*
MEL'ˇCUK I. (1988) *Dependency Syntax: Theory and Practice*. State University of New York Press.
MILLER G. A. & JOHNSON-LAIRD P. H. (1976) *Language and Perception*. Belknap Press of Harvard University Press.
MILLER G. A., LEACOCK C., TENGI R., BUNKER R. (1993) A semantic concordance, *Proceedings of the 3rd DARPA Workshop on Human Language Technology,* Plainsboro, New Jersey, March 1993, pp. 303-308.
MOORTGAT M. (1988) *Categorial investigations.* Dordrecht, Foris.
NG H.T. & LEE H.B. (1996) Integrating Multiple Knowledge Sources to disambiguate word sense : an examplar-based Approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* 24-27 june 1996, University of California, Santa Cruz, California, pp. 40-47.
OFLAZER K., BILGE S. & HAKKANI T. D. (2000) A Syntactic Annotation Scheme for Turkish. In *Proceedings of 10th International Conference on Turkish Linguistics (ICTL-2000*).
OOSTDIJ N. (1991) *Corpus Linguistics and the automatic Analysis of English. Amsterdam*, Rodopi.
OOSTDIJK N., de HAAN P. (1994) Clause patterns in Modern British English: A corpus-based (quantitative) studyICAME Journal N. 18, pp. 41-79
OSTLER N. & ATKINS B. T. S. (1992) *Predictable Meaning Shifts: Some Linguistic Properties of Lexical Implication Rules.*
PEREIRA F. C. N. (1985) A Structure-Sharing Representation for Unification-Based Grammar Formalisms. *ACL* 1985, pp. 137-144.
PEREZ I. (1994) Les agnosies auditives : une analyse fonctionnelle. In MCADAMS S. & BIGAND E. eds. *Penser les sons : psychologie cognitive de l'audition*. Paris, PUF, pp. 215 –248.
PIRON S. (2002a). Évolution sémantique des verbes de perception en français. Une approche lexicale, Communication donnée aux *XVIe Journées de Linguistique,* 14-15 mars 2002*,* AEDILL, Université

Laval, Québec, Canada.
PIRON S. (2002b). Les verbes de perception génériques en français. Une approche historique du lexique, Communication donnée à l'*ACFAS, Discipline Linguistique*, 14-15 mai 2002, Université Laval, Québec, Canada.
PIRON S. (2003) Description linguistique de la perception : le cas du verbe *entendre*. Communication donnée à l'*ACFAS,* Colloque « Description linguistique pour l'analyse automatique du français », 21-22 mai 2003, Université du Québec à Rimouski, Rimouski, Canada.
PIRON S. (2004, à paraître) Contraintes syntaxiques et préférences sélectionnelles du verbe *entendre*. In *Actes des Journées d'Analyse statistique de Données Textuelles*, mars 2004.
POLLARD C. J. & SAG I. A. (1988) *Information-Based Syntax and Semantics*. Chicago University Press, CSLI Lecture Notes.
POLLARD C. J. & SAG I. A. (1994) *Head-driven Phrase Structure Grammar*. Stanford University Press and University of Chicago Press, CSLI Lecture Notes.
PUSTEJOVSKY (1995) The Generative Lexicon. MIT Press.
PUSTEJOVSKY J. & BOGURAEV B. (1993) Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63, pp. 193-223.
PUSTEJOVSKY J. (1991) The generative lexicon. *Computational Linguistics*, 17(4).
PUSTEJOVSKY J. (1994) Linguistic constraints on type coercion. In P. St. Dizier and E. Viegas (ed) *Computational Lexical Semantics*. Cambridge University Press.
QUIRK R. (1970) Taking a deep smell. *Journal of Linguistics*, 6, pp. 119-124.
RANÇONNET & NICOT (1606) *Thresor de la langue française tant ancienne que moderne.*
RASTIER F. (2000) De la sémantique cognitive à la sémantique diachronique : les valeurs et l'évolution des classes lexicales. In *Théories contemporaines du changement sémantique*. Mémoires de la Société de linguistique de Paris, Peeters, pp. 135-164.
RAY E. T. (2003) *Learning XML*. Sebastopol, *O'Reilly. 2*e edition.
RESNIK P. (1997) Selectional Preferences and Sense Disambiguation. *Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?,* Washington, April 4-5 1997.
ROGERS A. (1971) Three Kinds of Physical Perception Verbs. *CLS 7,* pp. 206-223.
ROGERS A. (1972) Another Look at Flip perception Verbs, *CLS 8,* pp. 302-315.
ROMANSKI L. M. , TIAN B., FRITZ J., MISHKIN M., GOLDMAN-RAKIC P. S. & RAUSCHCKER J. P. (1999) Dual Streams of Auditory Afferent target Multiple Domains in the Primate Prefontal Cortex. In *nature Neuroscience,* 12, pp. 1131-1136.
ROSCH E. (1975). *Cognitive reference points*. Cognitive Psychology, 7, 532--547.
SAG I. A. & WASOW T. (1999) *Syntactic Theory. A Formal Introduction.* CSLI.
SANFILIPPO A. (1993) LKB encoding of lexical knowledge. In BRISCOE T., COPESTAKE A. & DE PAIVA V. (ed) *Default Inheritance within Unification-Based Approaches to the Lexicon*. Cambridge University Press.
SANFILIPPO A. (1995) Lexical polymorphism and word disambiguation. In *Working Notes of the AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*. Stanford University.
SANFILIPPO A., BENKEREMINI K. DWEHUS D. (1994) Virtual polysemy. In COLING94.
SHIEBER S. M. (1984) The Design of a Computer Language for Linguistic Information. In *Proceedings of COLING 84*, Association for Computational Linguistics, pp. 362-366
SHIEBER S. M. (1986) *An introduction to unification-based approaches to grammar*. CSLI Lecture Notes 4.
SHIEBER S. M. (1990) Les grammaires basées sur l'unification. In MILLER P. & TORRIS T. (eds) *Formalismes syntaxiques pour le traitement automatique du langage naturel*. Hermès, pp. 27-85.
SINCLAIR J. (1996) *Preliminary recommendations on corpus typology*, Rap. Tech., EAGLES, may 1996, CEE.
SINCLAIR J. (1997) Corpus Evidence in Language Description. In WICHMANN A. *et al.* (eds.) *Teaching and Language Corpora*. Longman, pp. 27-39.

SMITH G. W. (1991) *Computers and Human Language*. Oxford University Press.
SPERBERG-MCQUEEN C. M. & BURNARD L. (2002) *The Text Encoding Initiative Guidelines (P4)*. The TEI Consortium : The Association for Computers and the Humanities (ACH); The Association for Computational Linguistics (ACL); The Association for Literary and Linguistic Computing (ALLC).
SPERBERG-MCQUEEN C.M. & BURNARD L. ed. (2001) The TEI Consortium [ACH, ACL, ALLC] *TEI Guidelines for Electronic Text Encoding and Interchange (P4).* http://etext.lib.virginia.edu/teip4/
STEEDMAN M. (1996*) Surface structure and interpretation*. MIT Press.
STEEDMAN M. (2000) *Syntactic Process.* Bradford.
STUBBS M. (1996) *Text and Corpus Analysis : computer-assisted studies of language and culture*. Blackwell, Oxford.
TALMY, L. (1975). Semantics and syntax of motion. In *Syntax and Semantics*, volume 4. Academic Press, New York.
TALMY, L. (1988). Force dynamics in language and cognition. *Cognitive Science* 12(1), pp. 49-100.
TALMY, L. (2000). *Toward a cognitive semantics*. MIT Press.
VALLI A. (2000) Traitement automatique : introduction. In BILGER M. *Corpus. Méthodologie et applications linguistiques*, Paris, Champion, pp.77-81.
VAN DEVELDE R. (1977) Mistaken views of *see*. *Linguistic Inquiry,* 8, pp. 767-771.
VAN HALTEREN H. & OOSTDIJK N. (1993) Toward a syntactic database : the TOSCA analysis system. In Aarts J., de Haan P. & Oostdijk N. (eds) *English Language Corpora : Design, Analysis and Exploitation*. Amsterdam, Rodopi.
VAN VOORST J. (1992) The Aspectual Semantics of Psychological Verbs. *Linguistics and Philisophy,* 15, pp. 65-92.
VENDLER Z. (1967) *Linguistics and Philosophy*. Ithaca/NY, Cornell University Press.
VÉRONIS J. (2000) Annotation automatique de corpus : panorama et état de la technique. In Pierrel J.-M. *Ingénierie des langues*. Hermès, pp. 111-130.
VIBERG A. (1984) The Verbs of Perception : A Typological Study. In Butterworth B., Comrie B. & Dahl O. (eds) *Explanations of Language Universals.* Mouton, pp. 123-162.
Vossen P. ed. (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic*. Kluwer Academic Publishers.
WEIBE J., MAPLES J., DUAN L., BRUCE R. (1997) *Experience in WordNet sense tagging in the Wall Street Journal*, ACL-SIGLEX Workshop "Tagging text with lexical semantics : why, what and how ?", April 4-5 1997, Washington, D.C., pp. 8-11.