Prédiction du contexte droit des catégories prédicatives

Aurélie Merlo, Antonio Balvet, Rafael Marin, François Liger

Rapport final de la PTR *SéMoteur*
Laboratoire STL UMR 8163 (université Lille 3) – Ergonotics SAS

Prédiction du contexte droit des catégories prédicatives

Rapporteur : Aurélie Merlo (STL UMR 8163, Ergonotics SAS)

Coordinateurs : Antonio Balvet & Rafael Marin (STL UMR 8163) et François Liger (Ergonotics SAS)

octobre 2012

Sommaire

Introduction

Contexte

La PTR (Prestation Technologique Réseau) *SéMoteur* (Moteur Sémantique) implique l'entreprise Ergonotics SAS située à Villeneuve d'Ascq et le laboratoire de recherche STL UMR 8163 de l'université Lille 3.

L'entreprise Ergonotics est une jeune entreprise innovante créée en 2010. Elle développe des applications intuitives pour iPhone et iPad dont l'interaction homme-machine se fait en langage naturel. Pour ce faire, l'entreprise a élaboré un moteur d'analyse (breveté sous le nom d'*ActiveLinguistics*) prenant en charge des connaissances linguistiques et contextuelles[1]. Ce moteur permet aux utilisateurs d'interagir avec les applications de manière simple et intuitive. Pour exemple, l'application *convex*[2] pour iPhone/iPad commercialisée en 2011 est un convertisseur de mesure intelligent en langage naturel. L'utilisateur peut par exemple convertir rapidement et facilement en pouce une mesure en centimètre en tapant la requête *3 cm en pouces*. L'état actuel de la technologie de l'entreprise Ergonotics repose sur un moteur d'analyse linguistique permettant de reconnaître et d'extraire des informations comme des numéros de téléphone, des adresses postales, des dates ou encore des lieux et ce dans toutes ses applications.

L'interaction homme-machine en langage naturel offre des avantages que n'ont pas les autres types d'interaction (langages de commande, formulaires, menus, requêtes) :

- elle ne nécessite pas de connaissances particulières autre que celles sur notre langage ;
- de ce fait, un large public peut utiliser une application en langage naturel ;
- elle permet de faire des requêtes complexes avec de nombreux paramètres.

En revanche, en situation de mobilité, l'interaction homme-machine présente des inconvénients dus aux limitations physiques des nouveaux périphériques (écran de petite taille, clavier de petite taille, clavier virtuel, etc.) :

- une difficulté d'utilisation du clavier : la saisie au clavier est source d'erreurs de frappe même s'il existe des systèmes d'aide à la saisie (correcteur orthographique ou saisie intuitive) ;
- une lenteur d'utilisation du clavier par manque de concision de la langue ;
- des propositions beaucoup trop nombreuses faites par les systèmes d'aides à la saisie.

Par ailleurs, parmi les utilisateurs de smartphones et/ou des tablettes tactiles, il y a des personnes rencontrant des difficultés avec l'utilisation de tels périphériques comme les personnes âgées ou les personnes à mobilité réduite.

L'entreprise Ergonotics souhaite améliorer l'interaction homme-machine de ses applications iPhone et iPad en remédiant aux inconvénients cités ci-dessus. Pour cela, la solution envisagée est de proposer aux utilisateurs, parmi les préférences, une saisie intuitive capable de prédire les mots au fur et à mesure de la saisie. Les méthodes statistiques en saisie prédictive présentant des limites que nous montreront en première partie de ce rapport, l'entreprise Ergonotics envisage de développer un moteur de prédiction de mots à base de connaissances linguistiques. Cette solution

[1] Les informations contextuelles désignent les informations fournies à l'application soit par l'utilisateur lui-même (date de naissance, localisation, etc) soit par d'autres applications (par exemple l'heure et la date en cours fournies par les applications intégrées dans l'iPhone et l'iPad). L'utilisation de ces informations contextuelles permet par exemple à l'entreprise Ergonotics d'analyser la valeur de *demain* dans l'action *appeler Bob demain* et de replacer correctement l'action dans un agenda électronique.

[2] Site : http://www.convexapp.com/

devrait permettre (i) de réduire le nombre d'erreurs de frappe, (ii) d'augmenter la vitesse de saisie et (iii) de réduire le champ des possibles en ne proposant à l'utilisateur que les mots dont il aurait potentiellement besoin en fonction des mots déjà saisis (ex : pour les mots déjà saisis *je mange*, le champ des possibles proposé se limiterait à des noms renvoyant à des entités comestibles). Afin de déterminer quels types d'information linguistique sont nécessaires pour l'implémentation d'un tel moteur de prédiction, l'entreprise se tourne vers le laboratoire de recherche STL UMR 8163.

Le laboratoire STL UMR 8163 est composé de chercheurs en linguistique, philologie, philosophie et histoire des sciences. Les activités du laboratoire sont centrées autour de quatre axes. Le premier axe concerne la linguistique et s'intitule *Syntaxe, Interprétation, Lexique, Acquisition*. Cet axe est divisé en quatre thématiques : (i) lexique, (ii) syntaxe et sémantique de la phrase, du syntagme verbal et du syntagme nominal, (iii) discours, oralité(s), gestualité(s) et (iv) acquisition et didactique des langues. Le projet *SéMoteur* s'intègre particulièrement dans l'axe A thématique 1 du laboratoire STL UMR 8163 où le lexique est étudié sous ses aspects morphologique et sémantique.

Objectif scientifique

L'entreprise Ergonotics souhaite améliorer l'interaction en langage naturel de ses applications iPhone/iPad par l'implémentation d'un moteur de prédiction de mots à base de connaissances linguistiques.

Dans ce contexte, l'objectif scientifique du projet *SéMoteur* est de lancer une réflexion sur le(s) type(s) d'information nécessaire(s) pour améliorer l'interaction et évaluer les ressources linguistiques existantes pour la prédiction du contexte droit d'un mot.

Afin de réduire notre champ d'étude, car il n'est pas raisonnable d'étudier l'ensemble du lexique d'une langue, nous étudierons uniquement la prédiction du contexte droit des mots en français appartenant à une catégorie dite prédicative, autrement dit possédant au moins un complément : *manger* [*une pomme*]$_{comp.}$, *suppression* [*d'un poste*]$_{comp.}$, *libre* [*de partir*]$_{comp.}$

La notion de *prédicat* est à l'origine liée à la catégorie verbale. En syntaxe, Chomsky (1965) définit le prédicat comme étant le syntagme verbal directement dominé par *P*, *P* étant une proposition (Fig. 1 ci-dessous).

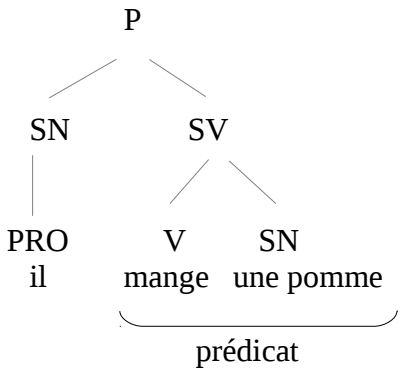


Fig. 1 – Représentation du prédicat selon Chomsky (1965)

Les mots en relation ici avec le verbe du prédicat sont appelés *arguments*. Selon les grammaires, les arguments peuvent être étudiés sur un même plan ou différenciés afin de distinguer les arguments dits « primaires » (obligatoires) des arguments dits « secondaires » (optionnels). Ainsi, dans l'exemple *Pierre construit une maison avec des cartes*, l'argument *maison* est primaire et

l'argument *avec des cartes* est secondaire car sa suppression n'entraîne pas une agrammaticalité de la proposition. Concernant l'argument secondaire, on parle aussi d'argument adjonctif ou d'adjoint.

Les notions de prédicat et d'argument ont été par la suite étendues à des catégories spécifiques de noms (noms déverbaux formés sur des verbes par exemple) et à certains adjectifs appelés réciproquement noms et adjectifs prédicatifs.

Concernant les verbes prédicatifs, nous nous préoccupons uniquement du contexte droit, autrement dit des arguments dits « internes » en position d'objet et non des arguments dits « externes » en position sujet. Concernant toutes les catégories prédicatives, nous nous limitons à la prédiction des arguments primaires (obligatoires) (ex : *Max parle à Léa [dans la cuisine]*$_{comp.\ secondaire}$). Dernier point de précision de notre champ d'étude, nous laissons de côté la prédiction des catégories mineures (les déterminants notamment) pour se centrer sur la prédiction des têtes syntaxiques au sein d'un constituant définies comme ce qui contrôle la distribution de ce même constituant. Dans l'exemple, *Max mange une pomme*, il s'agit de prédire *pomme* en contexte droit de *mange* et non pas le déterminant *une*. En ce sens, nous nous situons davantage du côté d'une prédiction de concepts que d'une prédiction de mots.

Annonce du plan

Dans une première partie, nous dressons un état des lieux de la prédiction de mots en présentant les différentes méthodes utilisées et leurs limites. Dans la seconde partie, nous étudions le contexte droit des verbes prédicatifs et nous évaluons les ressources attenantes sur une tâche de prédiction du contexte droit d'un verbe prédicatif. Nous faisons de même pour les noms prédicatifs dans la partie 3 et les adjectifs prédicatifs dans la partie 4.

Partie 1 : prédiction de mots

1. État de l'art

La conception d'une méthode de saisie efficace est devenue un objectif de recherche en interaction homme-machine et de multiples méthodes ont été proposées depuis dix ans. Nous proposons dans cette première partie une brève présentation d'une de ces méthodes : la méthode de saisie basée sur la prédiction de mots.

L'expression *prédiction de mots* désigne dans la littérature scientifique deux systèmes informatiques qu'il convient de distinguer. Le premier système est ce que l'on appelle plus communément la complétion de mot. Ce système, datant des années 80, permet d'obtenir une liste de mots probables au fur et à mesure que l'utilisateur entre une lettre. Le second système est appelé prédiction de mots et permet d'obtenir également une liste de mots probables en fonction cette fois des mots déjà entrés par l'utilisateur. Il existe également des approches de la prédiction de mots combinant ces deux systèmes.

Quelque soit le système utilisé ci-dessus, l'objectif de la prédiction de mots est le même. La prédiction de mots vise à proposer les mots susceptibles d'être sélectionnés par l'utilisateur en contexte droit en fonction de ce qui a déjà été sélectionné en contexte gauche. Il existe quelques cas particuliers de prédiction de mots consistant à prédire un ou plusieurs mots au milieu d'une phrase en s'aidant du contexte gauche et du contexte droit comme par exemple dans l'étude de van den Bosch (2006). Dans cette étude, il s'agit de prédire les mots *is tired* dans la phrase à trous *Alice was beginning to get very* (pour le contexte gauche) et *of sitting by her sister on the bank* (pour le contexte droit).

La prédiction de mots a donc pour objectif de réduire un maximum le champ des possibles du contexte droit en fonction du contexte gauche. Pour atteindre cet objectif, plusieurs méthodes ont été testées. Voici quelques-unes de ces méthodes relevées dans la littérature scientifique :

(i) des approches purement probabilistes qui utilisent des algorithmes comme l'algorithme de Markov (ou appelé également Prediction Suffix Trees (PST) (Pereira & *al.*)) ou utilisant la théorie *Latent Semantic Analysis (LSA)*(Wandmacher & Antoine 2006). Ces algorithmes calculent la probabilité de prédiction d'un mot à partir d'un corpus d'apprentissage de plusieurs millions de mots ;

(ii) des approches *n-grammes* qui estiment la fréquence d'une séquence de mots $word_1...word_N$ dans une langue donnée à partir d'un corpus d'apprentissage de plusieurs millions de mots (Nantais & *al*., 2001; Shein & *al*., 2001) ;

(iii) des approches hybrides mêlant *n-grammes* et connaissances syntaxico-sémantiques (Carlberger & *al.* 1997) ;

(iv) des approches qui extraient à partir d'un corpus les contraintes syntaxiques et/ou sémantiques qui régit la restriction des arguments au sein d'un prédicat donné (Sundarkantham & Mercy Shalinie) ;

(v) des approches purement linguistiques utilisant des lexiques de grandes tailles et dont chaque entrée contient des informations de fréquence et morpho-syntaxiques (systèmes : *Vitipi* (Boissière & Dours, 2000), *HandiAS* (Le Pévédic, 1997), *Kombe* (Pasero & Sabatier, 1995), *Sibylle* (Schadle, 2003)). D'autres systèmes de prédiction de mots prennent en compte également des informations sémantiques comme *Profet* décrit dans (Carlberger & *al*. 1997). Ce système est capable de prédire le contexte droit d'un mot à l'aide d'informations syntaxiques et d'un jeu de 4 étiquettes sémantiques

pour décrire les arguments : inanimé, animé, human, inanimé qui se comporte comme un humain.

Les méthodes (i), (ii), (iii) et (iv) sont des approches « corpus-based » ce qui pose deux problèmes pour la prédiction de mots : le corpus doit être de taille importante (plusieurs millions de mots) et équilibré. Péry-Woodley (1995 : 219) exprime la difficulté de constituer un corpus équilibré :

> En effet, la recherche de corpus équilibrés semble bien constituer une impasse : la notion d'équilibre s'apparente à celle de “langue générale”, et elle paraît tout aussi insaisissable. Elle suppose également une recherche irréaliste d'exhaustivité : le corpus équilibré est sans doute celui qui a “de tout un peu”, mais encore faudrait-il savoir ce qu'est “tout”, c'est-à-dire quelles sont les classes à représenter, – ce qui nécessite un modèle complet de la variation –, et avoir accès à des textes les représentant.

Notre approche de la prédiction du contexte droit des catégories prédicatives nous situe parmi les méthodes (v) purement linguistiques. Nous rappelons que notre objectif est de lancer une réflexion sur le(s) type(s) d'information linguistiques nécessaire(s) pour la prédiction des arguments des catégories prédicatives et évaluer les ressources linguistiques existantes pour cette tâche. Bien qu'il ne s'agisse pas d'une évaluation des différentes méthodes de prédiction du contexte droit, nous allons à présent décrire le protocole de test d'une méthode statistique de prédiction basée sur des n-grammes afin de la comparer à une méthode de prédiction de mots purement linguistique.

2. Évaluation d'une méthode statistique de la prédiction du contexte droit

- Définition de la méthode *n-grammes*

La méthode *n-grammes* permet de prédire le contexte droit d'un mot à partir d'une liste de *n-grammes* désignant une séquence de *n* mots, liste établie au préalable à partir d'un corpus. La méthode *n-grammes* peut également être utilisée pour la prédiction de lettres, un *n-gramme* désignant alors une suite de lettres consécutives.

Une liste de *n-grammes* est obtenue en déplaçant une fenêtre de *n* cases sur un texte ou un corpus de textes. Ce déplacement se fait de mot en mot et peut aller jusqu'à une fenêtre de cinq mots vers la droite. Par exemple, nous obtenons la série suivante de 5-grammes à partir de la phrase *« Max commence par remercier sa sœur dans son discours »* :

1. Max_commence_par_remercier_sa
2. commence_par_remercier_sa_soeur_
3. par_remercier_sa_soeur_dans
4. remercier_sa_soeur_dans_son
5. sa_soeur_dans_son_discours

À partir d'une telle liste de *n-grammes*, des calculs probabilistes sont appliqués pour déterminer quelle suite de mots est la plus probable après un mot donné.

- Protocole

Constitution du corpus de test

Pour l'évaluation de la méthode *n-grammes*, nous avons extrait du corpus de l'*Est Républicain* (année 2003)[3] 92 phrases contenant une occurrence d'une forme fléchie du verbe *commencer.* Le

[3] Corpus librement téléchargeable sur le site du Centre National de Ressources Textuelles et Lexicales : http://www.cnrtl.fr/corpus/estrepublicain/

choix de ce corpus s'explique par (i) le fait qu'il soit constitué d'articles journalistiques et de fait reflétant la langue courante, (ii) par sa structuration en XML facilitant l'extraction des occurrences et (iii) par le fait que ce soit un corpus récent.

Afin de comparer par la suite cette méthode avec une méthode purement linguistique, nous avons inséré dans le corpus 12 phrases agrammaticales et inacceptables du point de vue du sens (ex : **Max commence la peur*). Ces phrases sont indiquées dans le corpus par le signe *.

Corpus de phrases :

1. Je commence à 4 h 30.
2. Les conscrits d'Arçon ont commencé très fort mardi.
3. Pour bien commencer l'année, les musiciens, fidèles au rendez-vous, sillonnent les quartiers.
4. Une année s'achève, une autre commence.
5. J'ai commencé à avoir des contractions.
6. Sa voix, limpide et puissante, commence à marquer les esprits.
7. Une petite fête qui commencera par l'embrasement des sapins.
8. Des photos non officielles mais bien réelles commencent à circuler sur Internet.
9. L'année commence sur les chapeaux de roues.
10. Maurice commence sa vie professionnelle à la centrale de Nomexy.
11. Les travaux commencent par les logements pour la troupe.
12. Ses travaux devront donc commencer avant cette échéance.
13. C'est une grande aventure qui commence pour une équipe de passionnés.
14. Le fait de commencer à domicile peut être de bon augure pour la suite.
15. Damien a quitté la Maison familiale des Fins où il avait commencé une formation dans l'imprimerie.
16. Ils ont fait des études et commencent à travailler.
17. Le collège de Malzéville commence enfin les travaux de rénovation de locaux promis depuis longtemps.
18. Ils commencent à connaître les voisins.
19. Les 24 stagiaires ont commencé par des jeux autour du karaté.
20. Une nouvelle et sans doute grande aventure commence au royaume de l'ovalie.
21. Des travaux ont déjà commencé dans la cour d'honneur.
22. Cela commence par un test à l'effort.
23. Ils peuvent donc commencer à recruter.
24. Le football français commence doucement son petit marché de janvier.
25. Il commence par décrocher son combiné.
26. Un vœu-concours qui commence par ces mots tout simples.
27. Les souvenirs ont commencé à s'égrener.
28. Le public commence à s'habituer à cette nouvelle monnaie.
29. Les premières galettes de pétrole commencent à joncher les plages des Landes.
30. Cela commence à devenir une habitude.
31. Cette nouvelle édition commence par un rendez-vous devenu incontournable.
32. Les meilleures joueuses ont généralement commencé tôt.
33. J'ai commencé le billard très tard.
34. Beaucoup d'entre eux ont commencé la peinture seuls chez eux.
35. L'année officielle a commencé à Giromagny.
36. Les deux employés communaux ont commencé le déneigement des rues du village.
37. Nous avons commencé le salage.
38. Christine Cunin a commencé le dessin à 12 an.
39. Simone Thominot avait commencé l'arbre généalogique de sa famille.
40. Un vaste programme qui commencera avec la réfection de la fontaine.
41. Nos projets vont commencer à se concrétiser.
42. Il a commencé par rejoindre l'association.
43. D'où la décision de commencer les opérations de pompage du sous-sol.
44. Nous avons commencé une chorégraphie.
45. Le labo bisontin a commencé ses recherches en 2001.
46. Il a en effet commencé par recruter.
47. Coralie a commencé dans un cours privé.
48. Les flammes avaient déjà commencé d'attaquer la véranda en bois.
49. C'est là que l'affaire commence.
50. Sept ans est l'âge idéal pour commencer.
51. Je commence à désespérer.
52. Les Alsaciens commencent à entendre parler de notre ville.
53. Elle s'adjuge quatre victoires sur douze possibles à commencer par Raphaël Schild.
54. Les égoïstes bruiteurs commencent par respecter leurs voisins.
55. J'ai commencé par voir un camion dans le fossé.
56. Les membres du club de foot ont commencé la construction d'un bâtiment.
57. Les cours commencent par un échauffement.
58. Le tri sélectif commencera sur la commune.
59. Vincent commence un voyage initiatique.
60. Il commence à faire du judo.
61. On a commencé à déneiger une partie.

62. Le gaz a donc commencé à se répandre dans l'habitation.
63. L'école du Centre avait commencé un projet d'école sur trois ans.
64. Les enfants commencent par goûter.
65. J'ai commencé comme patrouilleur.
66. Il a commencé par demander une cigarette.
67. La procédure ne fait que commencer.
68. C'est un nouveau championnat qui commence.
69. Il faut donc commencer à gagner.
70. Les séances commencent par un échauffement spécifique.
71. Les patients commencent à affluer.
72. Denis Alliot a commencé cette réunion.
73. Les travaux ont déjà commencé.
74. On commence par une boule.
75. Il a commencé par créer le Shopi de Champenoux.
76. L'aventure commence assez mal.
77. Elle vient de commencer l'italien.
78. J'ai commencé la guitare à 8 ans.
79. Nicolas a commencé à travailler dans les hôpitaux.
80. Je commence à connaître un peu tout le monde.
81. *Max commence ranger.
82. *Max commence à la table.
83. *Max commence la peur.
84. *Max commence par terminer.
85. *Max commence à la lune.
86. *Max commence comme manger.
87. *Max commence de la lune.
88. *Max commence par la peur.
89. *Max commence le résultat.
90. *Max commence comme table.
91. *Max commence manger.
92. *Max commence comme rendez-vous.

Source de données

Le test de cette méthode se fait sur la prédiction du contexte droit du verbe *commencer* car ce verbe présente une alternance syntaxique. Il peut être employé dans un emploi absolu (ex : *Max commence*), dans une construction transitive directe autrement dit avec un objet direct (ex : *Max commence ses devoirs*) ou dans une construction intransitive autrement dit avec un objet indirect (ex : *Max commence à parler*).

La constitution d'une liste de *n-grammes* nécessite beaucoup de temps et de données. C'est pourquoi nous avons par conséquent utilisé une ressource déjà existante : Google Ngram Viewer[4]. Cette ressource comporte 500 milliards de mots extraits de la numérisation des livres recensés sur Google Books. Un site Web permet d'effectuer des requêtes pour obtenir des statistiques sur les mots (fréquence d'apparition sur une période donnée) et propose en téléchargement des fichiers contenant jusqu'à 5-grammes (suite de 5 mots). Nous avons sélectionné un fichier du site Google Ngram Viewer contenant des 5-grammes afin d'avoir une plus grande fenêtre du contexte droit. Ce fichier contient des données pour le français et ne concerne que les suites commençant par *co-*. Nous avons pu ainsi récupérer la liste des 5-grammes pour le verbe *commencer* contenant 2818 suites de mots différentes.

Préparation des données

Nous avons extrait de ce fichier les suites dont le premier mot est une occurrence d'une forme fléchie du verbe *commencer* (ex : commençons_à_connaître_un_certain). Afin de comparer plus facilement notre corpus de 92 phrases et la liste de 5-grammes pour le verbe *commencer*, nous avons lemmatisé ces deux fichiers. Par exemple, la suite de *n-grammes* « commençons_à_connaître_un_certain » ne permet pas une correspondance avec le contexte droit *à connaître* pour l'occurrence *je commence à connaître un peu tout le monde* de notre corpus car *commençons* et *commence* étant deux formes fléchies différentes de *commencer.* C'est pourquoi nous avons lemmatisé le corpus et les suites de *n-grammes* (ex :

[4] URL : http://books.google.com/ngrams

commencer_à_connaître_un_certain / *je **commencer à connaître** un peu tout le monde*).

- Résultats

La comparaison des 5-grammes pour le verbe *commencer* avec notre corpus de 92 phrases contenant des occurrences des formes fléchies de ce même verbe retourne 21 concordances (soit un taux de prédiction de mots de 22,83%). Les contextes droits concordants avec la liste des 5-grammes sont les suivants (indiqués en gras dans les phrases) :

1. J'ai **commencé à avoir** des contractions.
2. Des photos non officielles mais bien réelles **commencent à circuler** sur Internet.
3. Maurice **commence sa vie** professionnelle à la centrale de Nomexy.
4. Ils ont fait des études et **commencent à travailler.**
5. Ils **commencent à connaître** les voisins.
6. Un vœu-concours qui **commence par ces mots** tout simples.
7. Le public **commence à s'habituer** à cette nouvelle monnaie.
8. Cela **commence à devenir** une habitude.
9. Nos projets vont **commencer à se concrétiser.**
10. D'où la décision de **commencer les opérations** de pompage du sous-sol.
11. Le labo bisontin a **commencé ses recherches** en 2001.
12. Je **commence à désespérer.**
13. Les Alsaciens **commencent à entendre** parler de notre ville.
14. Les membres du club de foot ont **commencé la construction** d'un bâtiment.
15. Je **commence à connaître** un peu tout le monde.
16. Nicolas a **commencé à travailler** dans les hôpitaux.
17. Les patients **commencent à affluer.**
18. Il faut donc **commencer à gagner.**
19. Il a **commencé par demander** une cigarette.
20. Le gaz a donc **commencé à se répandre** dans l'habitation.
21. Il **commence à faire** du judo.

Cette évaluation de la méthode *n-grammes* pour la prédiction du contexte droit de *commencer* montre que cette méthode présente des limites : (i) nécessité de passer par une phase de lemmatisation, (ii) stricte concordance de mots implique de recenser toutes les combinaisons possibles et (iii) augmenter la taille du corpus pour obtenir un meilleur taux de prédiction. Or, malgré la taille importante des données de Google Ngram Viewer, nous obtenons seulement un taux de prédiction du contexte droit du verbe *commencer* de 22,83%. Par ailleurs, il n'est pas envisageable de stocker un tel corpus sur la mémoire d'un support mobile (smartphone ou tablette tactile).

Une autre possibilité serait d'envisager qu'il existe des contraintes linguistiques régissant le contexte droit du verbe *commencer* évitant ainsi de devoir lister toutes les combinaisons de mots possibles en contexte droit. Nous pouvons envisager une contrainte syntaxique telle que le verbe *commencer* est suivi d'un argument à l'infinitif introduit par la préposition *à* (ex : *Max commence à parler*). Cette approche purement linguistique permet alors de pallier les limites de l'approche *n-grammes*.

L'utilisation d'informations linguistiques dans la prédiction de mots présente d'autres avantages (Newelle & al. 1998 : 8-9) :

– la prédiction de phrases grammaticalement et sémantiquement correctes ;

– les informations linguistiques peuvent être aisément exploitées pour une seconde étape informatique après la prédiction comme une traduction automatique ;

– les informations linguistiques allègent la charge cognitive de l'utilisateur en lui proposant automatiquement une liste plus restreinte de mots probables en fonction du contexte syntaxique et sémantique en contexte gauche (MacKenzie, 2002; Garay-Vitoria & Abascal 2005).

Nous allons à présent évaluer cette approche purement linguistique consistant à prédire le contexte droit des catégories prédicatives à l'aide d'informations morpho-syntaxiques et

sémantiques. Pour cela, nous devons déterminer quelles sont les informations linguistiques qui doivent être prises en compte.

Partie 2 : prédiction du contexte droit des verbes prédicatifs

1. Définition d'un verbe prédicatif

- *Valence*

La notion de valence apparaît pour la première fois dans une étude de Tesnière (1959)[5] et est empruntée à la chimie. La valence caractérise le nombre d'arguments qu'un prédicat verbal doit avoir pour que la phrase dont il est le centre soit grammaticalement correcte. Le prédicat verbal « pleuvoir », par exemple, a pour valence 0, c'est-à-dire qu'il ne fait intervenir aucun argument : *il pleut*. En revanche, le prédicat verbal *manger* accepte 1 argument : *manger une souris*.

Dans le cadre de la prédiction du contexte droit, cette information nous permet déjà de mettre de côté l'ensemble des verbes qui ont une valence 0 comme le prédicat verbal pleuvoir.

- *Construction*

Lorsque l'on parle de la construction d'un verbe, il s'agit de déterminer s'il est de construction transitive ou intransitive. On appelle communément construction transitive une construction possédant au moins un argument et construction intransitive une construction ne possédant aucun argument. Concernant la construction transitive, il convient de distinguer la construction transitive directe avec la présence d'un argument en objet direct de la construction transitive indirecte possédant un argument introduit par une préposition.

(5) *Pierre mange* = construction intransitive
(6) *Pierre mange une pomme* = construction transitive directe
(7) *Pierre mange avec Marie* = construction transitive indirecte

La construction d'un verbe nous apporte une information supplémentaire dans le cas où un verbe est de construction transitive. En effet, cela nous permet de savoir si l'argument qui le suit est un objet direct (OD) ou un objet indirect (OI). Dans le cas d'une construction transitive indirecte, il est possible de connaître la ou les prépositions qui peuvent se trouvées après le verbe (*manger sur, manger avec*, etc.).

- *Cadre de sous-catégorisation* (CSC)

Un cadre ou schéma de sous-catégorisation (Messiant & *al.* 2008) correspond à la représentation syntaxique (catégorie et fonction) des arguments. Cela permet de savoir si l'argument d'un prédicat verbal en OD se réalise sous la forme d'un SN ou d'une infinitive par exemple. Le CSC est par conséquent plus précis en terme d'informations sur le contexte droit que le type de construction d'un verbe.

(8) *Pierre mange* -> [SUJ :SN]

[5] « On peut comparer le verbe à une sorte d'atome crochu susceptible d'exercer son attraction sur un nombre plus ou moins élevé d'actants, selon qu'il comporte un nombre plus ou moins élevé de crochets pour les maintenir dans sa dépendance. [...] Le nombre de crochets que présente un verbe et par conséquent le nombre d'actants qu'il est susceptible de régir, constitue ce que nous appellerons la valence d'un verbe » (Tesnière 1959 : 239).

(9) *Pierre mange une pomme* -> [SUJ : SN, OBJ : SN]
(10) *Pierre mange avec Marie* -> [SUJ:SN,P-OBJ:SP<avec+SN>]

Les informations contenues dans le CSC permettent d'exclure certaines agrammaticalité comme :
(11) **Pierre mange parler* / CSC [SUJ : SN, OBJ : SN]
L'objet du prédicat verbal *mange*r ne peut être un verbe à l'infinitif.

- *Structure argumentale*

À chaque argument du prédicat verbal est attribué un rôle thématique ou sémantique auquel correspond une projection syntaxique (voir aussi Hale & Kayser 1986a, 1986b ; Jackendoff 1983, 1987, 1990 ; Rappaport & Levin 1986 ; Grimshaw 1990). La réunion des informations sémantiques et syntaxiques des arguments (celles du CSC) correspond à la structure argumentale d'un prédicat verbal.

Les rôles thématiques ont été introduits par Levin (1993) puis repris par (Palmer & *al.* 2005) ce qui a donné lieu à deux ressources : PropBank (Palmer & *al.* 2005) et VerbNet (Kipper-Schuler. 2005). Le but de Levin est de classifier le lexique verbal anglais sous forme de classes et de sous-classes de verbes. Pour cela, elle part de l'hypothèse que le comportement syntaxique d'un verbe dépend de sa signification. Pour mieux représenter la signification d'un prédicat, elle a élaboré des rôles thématiques tels que pour le prédicat *manger* nous avons qqn (agent) manger qqch (patient).

Voici une description des rôles thématiques élaborés par Levin (1993) :

- Agent : ce qui est à l'origine d'un procès.
- Patient : ce qui est affecté par un procès.
- Thème : un participant à la phrase qui est à un endroit, ou qui va d'un endroit à un autre.
- Experiencer : un participant à la phrase qui est au courant de quelque chose (par exemple les sujets de verbes comme aimer, admirer).
- Stimulus : un évènement ou un objet qui apporte une réponse de type psychologique à un « Experiencer »
- Instrument : un objet qui engendre un changement d'état à quelque chose qui rentrerait au contact avec lui.
- Lieu : ce qui exprime un lieu.
- Source : ce qui caractérise le point de départ d'un mouvement ou d'un transfert.
- Goal : ce qui caractérise le point d'arrivée d'un mouvement ou d'un transfert.
- Récipient : ce qui est la cible d'un transfert.
- Benefactive : une entité qui bénéficie d'une action (en anglais, c'est le participant à l'alternance : « Benefactive alternation »)

Dans le cadre de notre objet concernant la prédiction du contexte droit, nous adressons plusieurs critiques à la notion de rôle thématique :

- les critères qui permettent d'attribuer un rôle thématique ne sont pas fixés
- il n'existe pas de liste finie de rôles thématiques
- un animal (le chat mange la souris) tout comme un humain (le chat a griffé l'enfant) peuvent être patient d'un procès selon le contexte. Par conséquent, dire que le verbe *manger* est suivi d'un argument dont le rôle thématique est *patient* retournerait les énoncés (1) l'enfant mange la souris et (2) ??la souris mange l'enfant. L'élimination de l'énoncé (2) implique une analyse linguistique complexe avant la prédiction du contexte droit du verbe *manger* pour déterminer le rôle thématique de l'argument en position sujet.

Les rôles sémantiques permettent une description des arguments d'un prédicat verbal de manière cognitive. Selon les tenants de la théorie de la sémantique cognitive (Croft & Cruse 2004), le langage n'est pas une faculté cognitive autonome, mais plutôt une structure conceptuelle. Les concepts sont associés par l'expérience, notamment RESTAURANT est relié aux concepts CLIENT, COMMANDE, ADDITION et NOURRITURE.
Nous pouvons alors imaginer un certain nombre de concepts autour du verbe *manger* tels que ALIMENT, LIEU, EVENEMENT, etc.

Il existe un dernier niveau de description sémantique des arguments qui consiste à leur attribuer des traits (+hum, +concret, +abstrait, etc.). Ces informations, contrairement aux rôles thématiques et rôles sémantiques, n'ont aucun rapport avec la relation que les arguments entretiennent avec leur prédicat. Ces traits sémantiques font référence aux caractéristiques extra-linguistiques des arguments.

2. Ressource existantes

Nous présentons les ressources existantes en français pour les verbes prédicatifs. Ces ressources contiennent pour la plupart des informations sur le cadre de sous-catégorisation accompagnées d'une caractérisation sémantique générale des arguments (+hum, -hum).

- Les verbes prédicatifs du *Lexique-Grammaire* (Gross 1975; Leclère 2002)

Les tables du LADL sont un lexique-grammaire établi par Maurice Gross et regroupant 6000 verbes répartis dans des tables construites d'après des similitudes de comportements des verbes qui les composent. Chaque table du lexique-grammaire contient un certain nombre de propriétés :

- les réalisations possibles des arguments ;
- les propriétés syntaxiques du verbe ou de ses arguments ;
- les sous-catégorisations alternatives ;
- les arguments ont certaines propriétés que Maurice Gross qualifie de « traits syntaxiques », (ce qui correspondrait plutôt à des traits sémantiques). Pour les arguments nominaux, le lexique définit si ce sont des arguments à trait humain ou non par exemple.

Ceci rejoint en partie la théorie de Beth Levin (Levin 93), selon laquelle les verbes partageant les mêmes traits syntaxiques et certains traits sémantiques.

N0 =: Nhum	N0 =: Nnr	N0 être V-n	N1 V	Ppv	Ppv =: se figé	Nég	<ENT>	N1 être V-n	N0 V Nhum sur ce point	N1 est Vpp W	N1 =: Nconc	N1 =: Nabs	<OPT>
+	+	-	-	<E>	-	-	abandonner	-	-	-	-	-	La chance abandonne Max
-	+	-	-	<E>	-	-	abasourdir	-	-	+	-	-	Le bruit a abasourdi Max
+	-	-	-	<E>	-	-	abattre	-	-	-	-	-	La police a abattu le truand
+	-	-	-	<E>	-	-	abattre	-	-	+	-	-	Le peuple abattra le tyran
~	~	~	~	<E>	-	-	aborder	~	~	~	~	~	Max a abordé un passant dans la rue
~	~	~	~	<E>	-	-	accabler	~	~	~	~	~	Le témoignage accable l'accusé
+	-	-	-	<E>	-	-	accolader	-	-	-	-	-	Le ministre a accoladé le héros
+	-	-	-	<E>	-	-	accompagner	-	-	+	-	-	Max accompagne Léa
+	-	-	-	<E>	-	-	accoster	-	-	-	-	-	Max a accosté une dame dans la rue

Fig. 2 – extrait d'une table du *Lexique-Grammaire des verbes prédicatifs*

- *Volem* (Saint-Dizier & *al.* 2002)

Volem (Saint-Dizier *et al.*, 2002) est une ressource multilingue (français-espagnol-catalan). Les entrées sont des verbes : la ressource décrit leur comportement syntaxique et sémantique à travers la description des arguments et des schémas de sous-catégorisation. Cette ressource décrit à l'heure actuelle 1700 verbes.

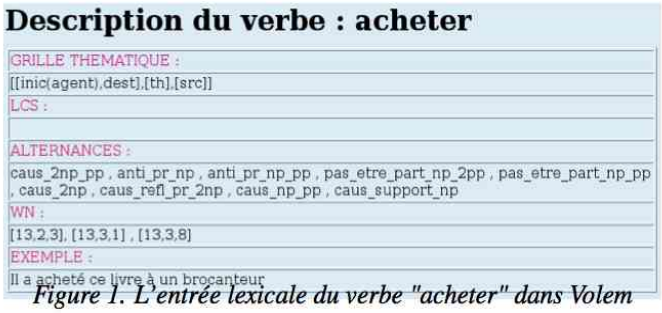


Figure 1. L'entrée lexicale du verbe "acheter" dans Volem

Fig. 3 – extrait de *Volem*

L'inconvénient de cette ressource pour notre objet d'étude de la prédiction du contexte droit d'un verbe prédicatif est l'absence de description précise des schémas syntaxiques que représentent les différentes alternances utilisées dans Volem.
Cette ressource n'est pas prise en compte dans notre évaluation de la méthode purement linguistique de la prédiction du contexte droit des catégories prédicatives car elle n'est pas disponible.

- Le *Lexique des Verbes Français* (Dubois & Dubois 1997)

Le *LVF,* « *Les Verbes Français* », est une ressource lexicale réalisée par Jean Dubois et Françoise Dubois-Charlier en 1997 dont l'objectif est de fournir une description linguistique des verbes basée sur l'adéquation entre schèmes syntaxiques et interprétation sémantique (Levin 1993).
Le *LVF* est composé de 25 609 entrées représentant les différents sens de 12 310 verbes. Il y a 4 188 verbes à plusieurs entrées et un verbe peut avoir jusqu'à 61 entrées (ex : *passer).* Une entrée est composée des onze champs suivants :

- MOT : entrée du verbe à l'infinitif
- DOMAINE : code donnant l'emploi principal (géologie, psychologie, . . .) et le niveau de langue (familier, vieux, littéraire, . . .)
- CLASSE : code définissant la classe syntactico-sémantique (appartenant à une hiérarchie)
- OPÉRATEUR : définition syntactico-semantique de l'entrée
- SENS : synonymes et définitions abrégées
- PHRASE : exemples d'utilisation de ce sens
- CONJUGAISON : codes permettant de conjuguer le verbe
- CONSTRUCTIONS : codes pour obtenir les schèmes de construction syntaxique
- DÉRIVATIONS : codes pour produire les adjectifs verbaux et les dérivés nominaux
- NOM : code pour produire le mot dont le verbe est dérivé
- LEXIQUE : code pour obtenir le type de dictionnaire où l'entrée est répertoriée

M	DOM	CLA	OPER	SENS	PHRASE	C	CONST	DER	N	L
amasser 01	OBJ	L3b	*lc.qp qc+pl e amas*	accumuler	On a~des documents,des livres. Les preuves s'a~contre P.	1aZ	T1801 P8001	-1- -D ---- --	3*	2
amasser 02	MON	L4b	*lc.qp arg e tas*	accumuler	On a de l'argent,de l'or.. On a~sans cesse.L'argent s'a~	1aZ	T1301 P3001	--- -- ---- --	-	5
amasser 03(s)	LOC	U1a	*(qn+pl)li.simul qp*	se grouper	La foule s'a~sur la place.	1aZ	P7001	--- -- ---- --	-	5

Fig. 4 – extrait du *Lexique des Verbes Français*

Une version xmlisée existe depuis 2004 développée par Hadouche et Lapalme (*à par.*). Cette version du *LVF* a pour objectif de rendre la ressource plus accessible, utilisable en TAL et extensible. Pour cela, ils ont informatisé la description des différents codes contenus dans la version papier du *LVF* sous la forme de fichiers XML.

- *LexValf* (Salkoff & Valli 2005)

La base de données *LEXique des VALences verbales du Français* (LexValf) réunit les verbes les plus fréquents de la langue française. Actuellement, cette base de données comporte 975 entrées ; à terme, la taille visée est de 6000 verbes.

Une entrée verbale contient une description syntaxique des arguments du prédicat verbal, elle est reliée à l'entrée correspondante dans la version xmlisée du *LVF* (2004), les relations de sélection lexicales entre le verbe et ses arguments et entre les arguments du verbes sont mises en évidence au moyen d'une annotation en traits sémantiques généraux (humain, non humain, concret, abstrait).
Les sources utilisées pour la constitution de cette ressource sont :

- des dictionnaires : *Grand Robert* (GR), *Petit Robert* (PR), et le *Trésor de la Langue Française* (TLF).
- La version XML du *LVF* de Dubois & Dubois : inventaire des construction possible pour une entrée verbale
- le Web : extraction de contextualisations, extraction de nouveau patrons de constructions syntaxiques

Les constructions possibles d'une entrée verbale sont extraites du Lexique-Grammaire des verbes prédicatifs. Les informations affichées sont : les patrons de compléments (SN, Vinf, etc.), les relations syntaxiques (Sujet, etc.) et des constantes grammaticales (Subj pour subjonctif...). Les patrons de compléments ont été réalisés à partir des tables du lexique-Grammaire. Exemple d'entrée de LexValf :

Verbe : commander
Emploi : commander 01 = diriger, être le chef [LVF]
Sous-catégorisation verbale : Aucune

Complément(s) = [] = [ ]
Restriction grammaticale: aucune restriction

N0[SN:N humain] commander
- Elle sait commander. [TLF]
- Ceux qui commandent ou administrent sont responsables. [TLF]

Fig. 5 – extrait de *LexValf*

Cette ressource n'est pas prise en compte dans notre évaluation de la méthode purement linguistique de la prédiction du contexte droit des catégories prédicatives car elle n'est pas disponible.

- *DicoValence* (van den Eynde & Mertens 2006)

DicoValence est un lexique décrivant la structure argumentale de plus de 3 700 verbes simples du français, élaboré dans le cadre de l'approche pronominale (Blanche-Benveniste & *al*. : 1984).

Dans cette ressource, la description de la structure argumentale est assez fine : aux 3 700 verbes correspondent plus de 8 000 cadres valenciels. Elle se fait de la façon suivante :

> [D]'abord, pour chaque place de valence (appelée « paradigme ») le dictionnaire précise le paradigme de pronoms qui y est associé et qui couvre « en intention » les lexicalisations possibles ; ensuite, la délimitation d'un cadre de valence, appelé « formulation », repose non seulement sur la configuration (nombre, nature, caractère facultatif, composition) de ces paradigmes pronominaux, mais aussi sur les autres propriétés de construction associées à cette configuration, comme les « reformulations » passives.

Les paradigmes les plus courants et les plus utilisés dans un schéma valenciel sont : P0 (paradigme 0), qui correspond *grosso modo* à l'argument externe ; P1 (≈ Objet Direct) ; P2 (≈ Objet Indirect, les formes non clitiques présentant la préposition *à*).

D'autres paradigmes sont le paradigme locatif (PL), le paradigme de manière (PM), le paradigme de temps (PT), ou le paradigme de quantité (PQ).

Le tableau suivant illustre l'emploi de ces paradigmes :

Verbe	type	SA	Exemple
circuler	intransitif	P0	*l'eau froide circule dans ces tubes*
disparaître	inaccusatif	P0 (PDL)	*son ami a disparu de la cabine téléphonique*
construire	transitif	P0 (P1)	*l'hiver, ils construisent des cabanes*
contribuer	transitif ind.	P0 (P2)	*son action contribue à la destruction des structures établies*
concéder	(di)transitif	P0 P1 (P2)	*on ne lui a concédé aucun droit*

Tableau 1 – informations contenues dans *DicoValence*

De la même manière que le *Lexique-Grammaire des verbes prédicatifs*, *DicoValence* contient des informations sémantiques sur les arguments sous la forme de traits sémantiques : /+Hum/, /+Nhum/, /+Abs/, etc.

- *Lefff* (Sagot & *al.* 2006)

Le *Lefff* (Lexique des Formes Fléchies du Français) est un lexique à large couverture. Chaque entrée se voit associé un cadre de sous-catégorisation (arg) et une liste des redistributions possibles (%) à partir de ce cadre :

clarifier1 Lemma;v;<arg0 :Suj:cln|scompl|sinf|sn,arg1 :Obj:(cla|scompl|sn)>; %ppp_employé_comme_adj,%actif, %se_moyen_impersonnel, %passif_impersonnel,%passif

Les fonctions syntaxiques sont définies dans le *Lefff* par des critères proches de ceux de DicoValence (Van den Eynde & Mertens 2006) : Suj (sujet), Obj (objet direct), Objà (objet indirect introduit canoniquement par la préposition *à)*, Objde (objet indirect introduit canoniquement par la

préposition *de)*, Loc (locatif), Dloc (délocatif), Att (attribut), Obl ou Obl2 (autres arguments obliques). Pour plus d'informations sur les critères définitoires de ces fonctions voir (Sagot & Danlos 2007).

Chaque fonction syntaxique peut être réalisée par trois types de réalisations : *pronoms clitiques*, *syntagme direct* (syntagme nominal (sn), adjectival (sa), infinitif (sinf), phrastique fini (scompl), interrogative indirecte (qcompl)) et *syntagme prépositionnel* (syntagme direct précédé d'une préposition, comme de-sn, à-sinf ou pour-sa.

Des informations syntaxiques complémentaires (contrôle, mode des complétives, etc.) sont notées par des *macros* (@CtrlSujObj, @ComplSubj, etc.).

- *SynLex* (Gardent & *al.* 2006)

SynLex est un lexique obtenu à partir des tables du lexique-grammaire. Les étapes qui ont permis sa création sont :

1. représentation manuelle du contenu des tables sous la forme d'un graphe dont les nœuds contiennent des conditions et des pointeurs vers le contenu des colonnes des tables du lexique-grammaire ;
2. production automatique à l'aide du graphe d'un lexique syntaxique représentant le contenu des tables ;
3. simplification du lexique produit pour ne contenir que les informations habituellement présentes dans un lexique syntaxique.

Une entrée de SynLex comporte un verbe, une liste d'arguments syntaxiques ayant un rôle sémantique, une liste optionnelle d'*associés* (arguments régis par le verbe mais sans rôle sémantique, *il pleut*), d'une liste de *macros* (informations supplémentaires sur les propriétés syntaxiques du verbe. Ce lexique contient 5244 verbes et 19127 entrées (paires verbe/cadre de sous-catégorisation). On recense 726 cadres de sous-catégorisation (avec les associés) et 538 (sans les associés).

Cette ressource n'est pas prise en compte dans notre évaluation de la méthode purement linguistique de la prédiction du contexte droit des catégories prédicatives car elle n'est pas disponible.

- *LexSchem* (Messiant & *al.*, 2008)

Le lexique *LexSchem* est un lexique de verbes français construit automatiquement à partir d'un corpus à large couverture (*Le Monde*). Le lexique contient 11 149 entrées lexicales qui se distinguent par leur cadre de sous-catégorisation. Le lexique comporte au totat 3268 lemmes verbaux et 336 cadre de sous-catégorisation. Chaque entrée est caractérisée par 7 types d'information :

- NUM: identifiant de l'entrée
- SUBCAT: verbe cible et cadre de sous-catégorisation de l'entrée lexicale
- VERB: le verbe
- SCF: le cadre de sous-catégorisation
- COUNT: le nombre d'occurrences trouvées dans le corpus pour le verbe et le cadre de sous-catégorisation en question
- RELFREQ: fréquence relative du cadre de sous-catégorisation
- EXAMPLES: occurrences relevées dans le corpus

```
:NUM:       05204
:SUBCAT:    s'abattre : SP[sur+SN]
:VERB:      S'ABATTRE+s'abattre
:SCF:       SP[sur+SN]
:COUNT:     420
:RELFREQ:   0.882
:EXAMPLE:   25458;25459;25460;25461;25462
```

Fig. 6 – extrait de *LexSchem*

- *TreeLex* (Kupść & Abeillé 2008)

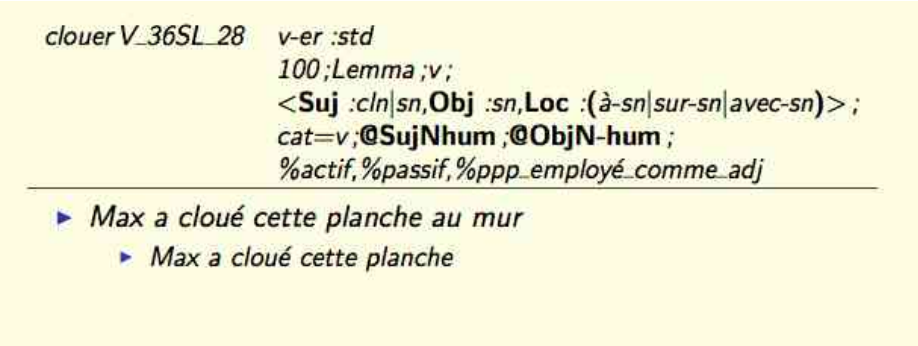
Le lexique *TreeLex* est un lexique français obtenu automatiquement et comporte des informations sur les cadres de sous-catégorisation de verbes et d'adjectifs prédicatifs extraits à partir du corpus annoté syntaxiquement le French TreeBank. Le lexique *TreeLex* se présente donc sous la forme de deux ressources distinctes, l'une pour les verbes l'autre pour les adjectifs. La ressource concernant les verbes contient 2000 verbes accompagnés d'une description de leur cadre de sous-catégorisation (nommé valence frames). Il y a 180 cadres de sous-catégorisation différents et une moyenne de 2,09 cadres de sous-catégorisation par verbe. Concernant la ressource pour les adjectifs, elle comporte 2153 adjectifs qualificatifs et 16 410 occurrences. 41 cadres de sous-catégorisation sont présents dans la ressource. La plupart de ces adjectifs apparaissent avec un cadre de sous-catégorisation simple (NP sujet) (1849 adjectifs). Seulement 304 adjectifs ont un cadre de sous-catégorisation ”complexe” (nommés les adjectifs ”intéressants”)

Les informations concernant les arguments (pour les verbes et les adjectifs) sont de deux types : catégorie syntaxique et fonction syntaxique.

- *LGLex* (Tolone 2011)

LGLex est un lexique obtenu à partir des tables du lexique-grammaire des verbes prédicatifs pour une utilisation en TAL. Une entrée du lexique contient 5 champs :

- ID : identifiant de l'entrée
- lexical-info : lemme et ses informations lexicales
- args : arguments et leurs distributions accompagnées de d'autres informations
- all-constructions : liste des constructions acceptées
- example : occurrences illustrant l'entrée



Adaptation du lexique au format ALEXINA pour une fusion avec *Lefff* :

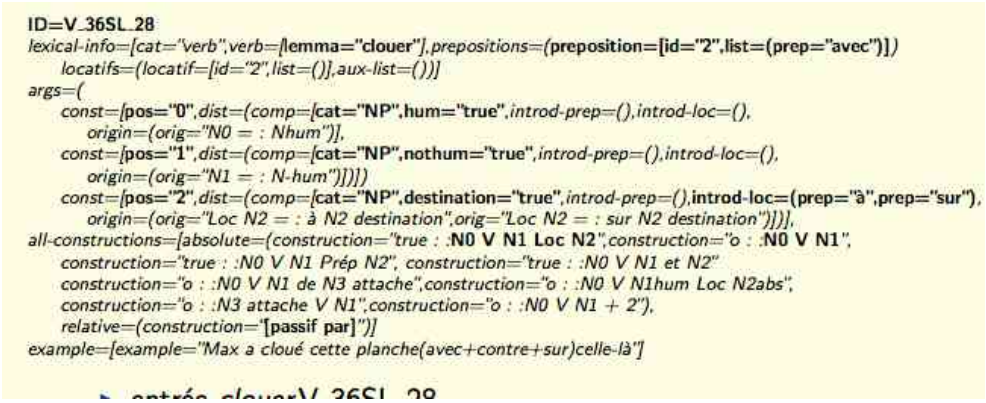


Fig. 7 – extrait de *LGLex*

Avant de passer à l'évaluation d'une méthode purement linguistique de la prédiction du contexte droit des verbes prédicatifs, voici un tableau récapitulatif des caractéristiques des ressources que nous venons de présentées et que nous allons évaluer :

Ressource	**Taille**	**Prévu pour une utilisation informatique**	**Informations syntaxiques**	**Informations sémantiques**
LVF	12310	oui	type de construction	classes sémantiques
Lefff	6818	oui	cadre de sous-catégorisation	traits sémantiques généraux
LGLex	5723	oui	cadre de sous-catégorisation	traits sémantiques généraux
DicoValence	3700	non	cadre de sous-catégorisation	traits sémantiques généraux
LexSchem	3268	oui	cadre de sous-catégorisation	-
TreeLex	2000	oui	cadre de sous-catégorisation	-

Tableau 2 – récapitulatif des ressources évaluées pour la prédiction du contexte droit des verbes prédicatifs

3. Évaluation des ressources

Nous venons de présenter les ressources existantes en français pouvant être exploitables pour la prédiction du contexte droit des verbes prédicatifs. Il s'agit à présenter d'évaluer leur efficacité.

Notre démarche consiste à sélectionner les informations pertinentes pour notre objet d'étude dans chacune de ces ressources pour le verbe *commencer*. Le choix de ces verbes s'est fait selon leur alternance de construction (transitif direct/indirect, intransitif). Ces informations sont utilisées sous forme de graphes de reconnaissance UNITEX qui seront par la suite appliquées sur un corpus test.

Concernant le Lexique-Grammaire, nous préférons exploiter les informations contenues dans LGLex (nous rappelons que cette ressource est une adaptation du Lexique-Grammaire sous forme XML) car les informations y sont plus clairement structurées.

- **Protocole**

Constitution du corpus de test

L'évaluation des ressources présentées précédemment porte sur la prédiction du contexte droit du verbe *commencer*. Le choix de ce verbe s'explique par son alternance syntaxique. Ce verbe peut être employé dans un emploi absolu (ex : *Max commence*), dans une construction transitive directe autrement dit avec un objet direct (ex : *Max commence ses devoirs*) ou dans une construction intransitive autrement dit avec un objet indirect (ex : *Max commence à parler*). Le fait que nous ayons déjà constitué un corpus d'occurrences de ce verbe pour l'évaluation de la méthode *n-grammes* dans la première partie justifie également notre choix.

Notre corpus est par conséquent constitué de 92 phrases contenant des occurrences de formes fléchies du verbe *commencer* dont 12 phrases agrammaticales et inacceptables du point de vue du sens.

Préparation des données

La seconde étape de notre protocole consiste en l'adaptation des données contenues dans les ressources présentées précédemment sous forme de graphes de reconnaissance Unitex.

Il n'existe pas de programme libre de prédiction du contexte droit. Dans le temps qui nous était imparti, le développement d'un tel programme était impossible. C'est pourquoi, nous avons utilisé l'outil linguistique Unitex[6] afin de simuler une prédiction du contexte droit. Unitex est un logiciel libre permettant le traitement de corpus pour construire ou gérer des ressources linguistiques. Le traitement des corpus se fait à l'aide de grammaires et de dictionnaires que l'on applique sur les textes.

Pour chaque ressource vue dans la section précédente, nous avons extrait les informations linguistiques pertinentes à la prédiction du contexte droit du verbe *commencer.* Par exemple, nous avons extrait les informations suivantes de *Lefff* :

pred="commencer_____1<Suj:(cln|scompl|sinf|sn),Obj:(cla|de-sinf|sn|‡-sinf)>"

Le lexique *Lefff* nous indique que le prédicat *commencer* est constitué de deux arguments : un argument externe remplissant la fonction sujet et un argument interne remplissant la fonction objet. Dans le cadre de la prédiction du contexte droit, l'objet nous intéresse particulièrement. Cette ressource nous indique que l'objet du verbe *commencer* peut être un syntagme nominal (*sn*), un infinitif introduit par la préposition *de* (*de-sinf*) ou un pronom clitique *(cla*).

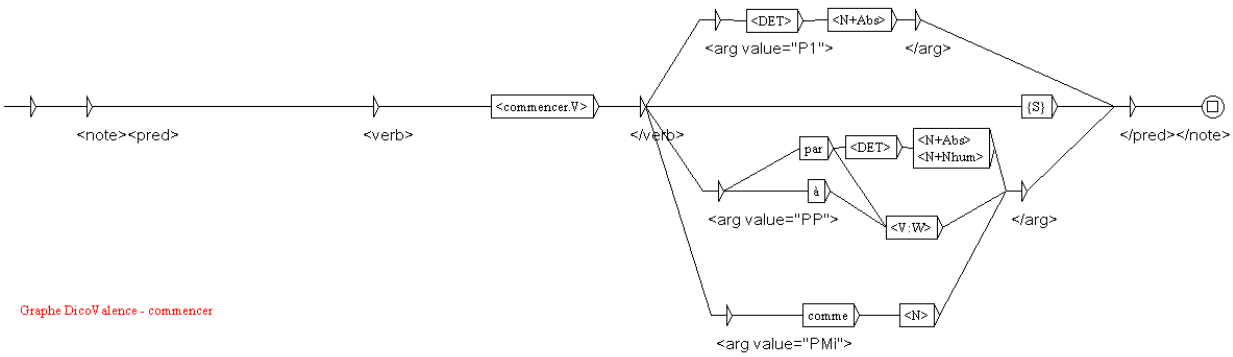
Ces informations ont été adaptées sous forme de graphes Unitex permettant la reconnaissance

[6] URL : http://www-igm.univ-mlv.fr/~unitex/

dans notre corpus des contextes droits du verbe *commencer*.

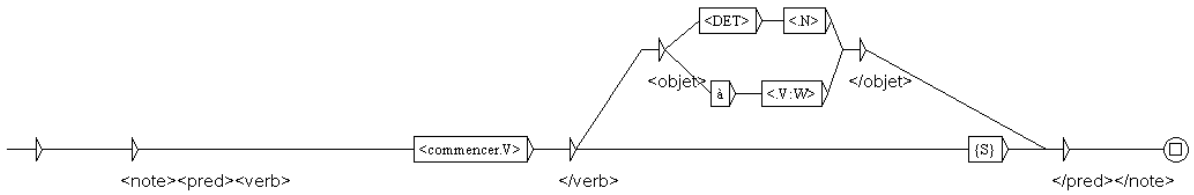
Élaboration des graphes de reconnaissance :

- **Graphe de reconnaissance pour *DicoValence***



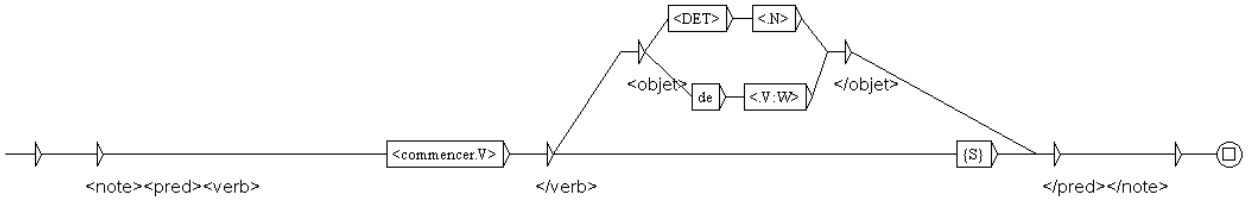
Graphe DicoValence - commencer

- **Graphe de reconnaissance pour *TreeLex***



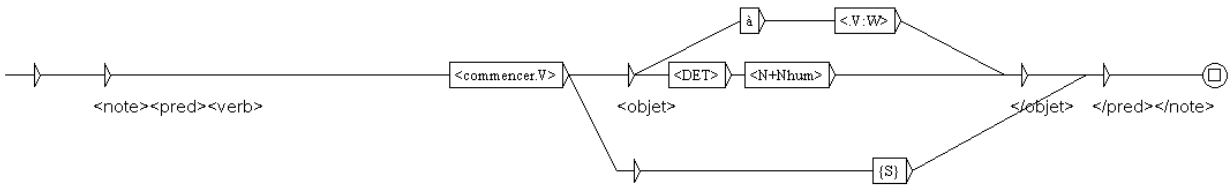
Graphe TreeLex - commencer

- **Graphe de reconnaissance pour *Lefff***



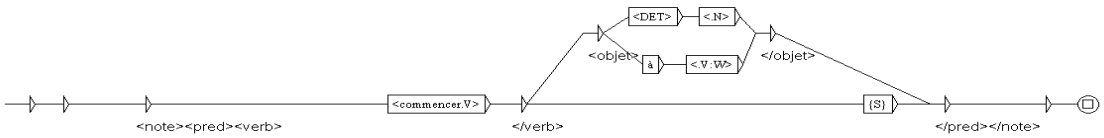
Graphe LefffF - commencer

– **Graphe de reconnaissance pour *LVF***



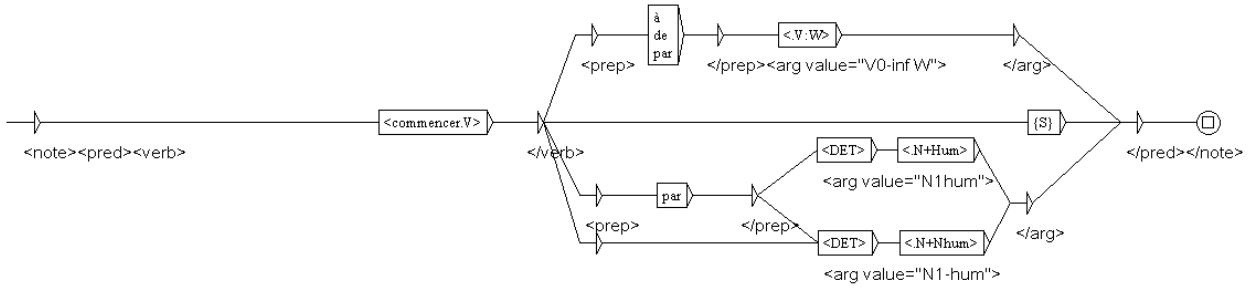
Graphe Lexique des Verbes Français - commencer

– **Graphe de reconnaissance pour *LexSchem***



Graphe LexSchem - commencer

– **Graphe de reconnaissance pour *LGLex***



Graphe LGLex - commencer

■ **Résultats**

L'application des graphes de reconnaissance sur notre corpus de phrases retourne les résultats suivants :

Ressource	**Nombre de contextes droits incorrects reconnus (/12)**	**Nombre de contextes droits reconnus (/92)**
DicoValence	12	56 (61%)
LGLex	5	52 (56%)
LexSchem	2	34 (37%)
TreeLex	2	34 (37%)
LVF	3	34 (37%)
Lefff	2	18 (19%)

Tableau 3 - résultats de la prédiction du contexte droit de *commencer* à base de connaissances syntaxiques

Ressource	**Nombre de contextes droits incorrects reconnus (/12)**	**Nombre de contextes droits reconnus (/92)**
DicoValence	7	51(55%)
LGLex	4	51 (55%)
LexSchem	2	34 (37%)
TreeLex	2	34 (37%)

LVF	2	33 (37%)
Lefff	2	18 (19%)

Tableau 4 - résultats de la prédiction du contexte droit de *commencer* à base de connaissances syntaxiques et sémantiques

Ces résultats sur l'évaluation des ressources existantes pour la prédiction du contexte droit des verbes prédicatifs apportent quatre informations intéressantes :

– l'intégration des connaissances syntaxiques dans la prédiction du contexte droit du verbe *commencer* permet à elle-seule d'améliorer le taux de prédiction par rapport à la méthode *n-grammes* testée dans la partie précédente (22,83%). En effet, mis à part la ressource *Lefff,* les taux de prédiction du tableau 3 sont tous supérieurs. Par conséquent, connaître la nature grammaticale des arguments d'un verbe prédicatif permet déjà de réaliser une meilleure prédiction.

– l'intégration des connaissances sémantiques dans la prédiction du contexte droit du verbe *commencer* fait diminuer le taux de prédiction des ressources contenant des informations sémantiques (voir dans le tableau 4 *DicoValence, LGLex, LVF*).

– d'un point de vue purement quantitatif, la ressource qui offre le meilleur taux de prédiction du contexte droit du verbe *commencer* est *DicoValence* : 61% à partir de connaissances syntaxiques et 55% à partir de connaissances syntaxiques et sémantiques. En revanche, la ressource *Lefff* a le taux de prédiction le moins élevé dans les deux tableaux. Cela s'explique par une description des cadres de sous-catégorisation possibles pour le verbe *commencer* plus complète pour *DicoValence* que pour *Lefff.* Dans la ressource *DicoValence*, le verbe possède 5 cadres de sous-catégorisation différents alors que *Lefff* n'en décrit que 3.

– d'un point de vue qualitatif, la ressource qui offre le meilleur taux de prédiction du contexte droit du verbe *commencer* est *LGLex.* En effet, il faut regarder le rapport nombre de contextes droits incorrects reconnus sur le nombre de contextes droits reconnus au total, soit pour *LGLex* 4 contextes droits incorrects reconnus sur 51 contextes droits reconnus au total (8%). La ressource *DicoValence* permet la prédiction de 7 contextes droits inconnus sur 51 contextes droits reconnus au total (14%).

Dans l'objectif de réduire le champ des possibles concernant le contexte droit des verbes prédicatifs afin d'améliorer l'interaction homme-machine en contexte de mobilité, nous considérons que la ressource la plus adaptée est celle qui permet de retourner le plus grand nombre de résultats tout en limitant la sur-prédiction. La ressource qui correspond à cette attente est la ressource *LGLex.*

Partie 3 : prédiction du contexte droit des noms prédicatifs

1. Définition d'un nom prédicatif

Si on prend l'exemple des verbes prédicatifs que nous venons d'étudier, nous pouvons donner une définition générale du nom prédicatif comme étant un nom possédant une structure argumentale. Mais que doit-on considérer comme argument lorsqu'il s'agit d'un nom ? Quels sont les noms concernés ?

Bien qu'il existe de nombreuses références sur la structure argumentale des noms (voir entre autres, Badia & Colominas (1997), Bierwisch (1990-1991), Grimshaw & Williams (1993), Ingria & *al.* (1993), Levin & Rappaport (1988), Gross (1989), Nunes (1993), Rappaport (1983), Roeper (1993), van Hout (1991), Williams (1981 et 1987)), il n'existe pas de consensus sur la définition d'un nom prédicatif dans la littérature scientifique. Par exemple, Badia (1994 : 63) avance que la majorité des noms prédicatifs sont formés à partir d'une base verbale ou adjectivale. En revanche, pour Gross (1989), la définition d'un nom prédicatif ne se résume pas à un lien morphologique avec une base verbale ou adjectivale mais doit prendre en compte également le fait qu'un nom prédicatif a des arguments.

L'étude centrale sur les noms prédicatifs est celle réalisée par Grimshaw (1990) qui distingue le participant sémantique d'un prédicat nominal et l'argument syntaxique d'un nom. La structure argumentale d'un nom prédicatif est vue comme une représentation lexico-syntaxique. Il existe une *Structure Lexico-Conceptuelle* (SLC) représentant les actants sémantiques du prédicat nominal et une *Structure Profonde* purement syntaxique. Selon Grimshaw (1990), tous les noms possèdent une SLC seulement tous les noms ne possèdent pas une structure argumentale (ils n'ont pas nécessairement d'arguments syntaxiques). Elle nomme *noms d'évènements complexe* les noms possédant une SLC et une structure profonde (ce qui constitue une structure argumentale) et *noms de résultat* les noms ne possédant qu'une SLC. Les noms de résultats sont définis comme désignant le résultat d'un procès. Par conséquent, Grimshaw admet que les noms de résultat entraînent l'existence de participants dans un contexte donné. Mais, selon Grimshaw, ces participants jouent un rôle dans la SLC et non pas dans la structure argumentale. D'autres noms se comportent comme les noms de résultat selon l'auteure : les noms avec complément phrastique n'ont pas de structure argumentale. Pour Grimshaw, des noms comme *justification* se comportent comme des noms de résultat, c'est-à-dire qu'ils n'ont pas de structure argumentale (ex : *la justification que les dépenses ont augmenté ne me convainc pas*). C'est seulement lorsqu'il y a obligation d'exprimer les arguments que l'on peut parler de structure argumentale.

D'autres approches théoriques avancent l'hypothèse que les noms de résultat peuvent avoir des arguments. Ainsi, pour Pustejovsky (1995), un nom de résultat en anglais comme *construction* ou *arrival* est une unité lexicale ayant un *dotted type* (type complexe) formé par « procès-état ». C'est pourquoi, il est possible de relever des occurrences de *construction* pouvant désigner uniquement le procès (*The construction was arduous and tedious*) ou uniquement le résultat (*The construction is standing on the next street*). Bierwisch (1990-1991) ajoute que d'autres noms peuvent faire référence à un procès ou un objet comme *book* dans la phrase *the book is entertaining, inexpensive and easy to take along.*

Une corrélation entre la nature sémantique du nom et la structure argumentale a été également mise en évidence dans la littérature. Il semble que l'on assimile de préférence noms abstraits et structure argumentale. Demonte (1989) signale que des noms concrets comme *table* ou *arbre* n'assignent pas de rôle sémantique puisqu'ils sont privés d'une structure événementielle. Gross (1989) identifie également nom prédicatif et nom abstrait car un nom concret est « inerte et n'est pas susceptible de recevoir aucune indication de temps et de personne, bref qu'il ne s'agit pas d'un

événement ou d'un procès » (G. Gross 1989 : 22).

Enfin, la définition d'un nom prédicatif peut prendre en compte une dimension aspectuelle. Étant donné que la structure argumentale a deux dimensions, sémantique et aspectuelle (Grimshaw 1990), seuls les noms dont le sens a une dimension aspectuelle ont une structure argumentale (Tenny 1992, 1994 ; van Hout 1991).

2. Ressources existantes en français

Les ressources qui existent actuellement pour les noms prédicatifs sont bien moins nombreuses que les ressources recensées précédemment pour les verbes prédicatifs. Pour le français, nous recensons 4 ressources : le *Lexique-Grammaire des noms prédicatifs* de M. Gross, *LGLex*, *Lefff* et le lexique *Nomage*.

Les ressources *LGLex* et *Lefff* ayant été déjà décrites dans la partie 1, nous ne reviendrons pas dessus.

- Les noms prédicatifs du *Lexique-Grammaire* (Leclère 2002)

Nous ne reviendrons pas plus en détail sur cette ressource déjà décrite dans la partie 1 précédente. Cette ressource contient 59 tables de noms prédicatifs (noms avec argument(s) étudiés avec leur verbe support[7]). Les informations qui nous intéressent particulièrement dans la prédiction du contexte droit d'un nom prédicatif sont :

Informations contenues dans les tables des noms prédicatifs	**Exemple**
nom prédicatif	*affection*
préposition qui entre dans la construction du nom prédicatif (Prép 1)	pour
description sémantique de l'argument N1 (traits sémantiques généraux : humain/non humain)	N1 : Nhum
structure argumentale du nom prédicatif	GN : le N de N0 Prép N1 (l'affection de Max pour ses enfants)
exemple	*Max a de l'affection pour ses enfants*

Tableau 5 – informations contenues dans le *Lexique-Grammaire des noms prédicatifs*

- Lexique *Nomage* (Balvet & *al.* 2012)

Le lexique NOMAGE est issu d'un projet de recherche ANR jeunes chercheurs de même nom dont l'objectif était d'étudier les noms déverbaux afin de déterminer les propriétés qu'ils héritent de leur base verbale. Les propriétés en question sont la structure argumentale et la classe aspectuelle.

Des occurrences de noms déverbaux ont été extraits à partir d'un corpus annoté morpho-syntaxiquement (*French TreeBank*, Abeillé 2003). Une liste de suffixes identifiés comme entrant dans la formation de noms déverbaux ont permis d'extraire automatiquement ces occurrences (*-eur, -ment, -ion, -ure, -age, -ance/ence, -éé* et *-ade*). Puis, la liste Verbaction a permis de nettoyer les occurrences des « faux » noms déverbaux (ex : *pommade*). Une fois le corpus NOMAGE créé, une

[7] Un verbe support est un verbe « qui a comme objet (direct ou indirect) un nom prédicatif (N_{pred}) dénotant une éventualité, comme sujet un participant à cette éventualité – en gros, le participant qui est le sujet du verbe morphologiquement associé au N_{pred} s'il existe – et comme objet oblique éventuel l'autre participant s'il y en a un » (Danlos 2009).

série de 10 tests a été appliquée sur les noms déverbaux afin de déterminer leur classe aspectuelle

Il existe d'autres ressources sur les noms prédicatifs pour d'autres langues : *NomBank* et *NomLex* pour l'anglais, *AnCora* pour l'espagnol.

3. Ressources existantes dans d'autres langues

- *NomLex* (Macleod & *al.* 1998)

NomLex[8] (NOMinalization LEXicon) est un lexique, un dictionnaire de nominalisations anglaises développé dans le cadre du Proteus Project par l'Université de New York sous la direction de Catherine MacLeod. Le but de ce projet vise à déterminer quels sont les compléments autorisés pour une nominalisation et à mettre en relation les compléments nominaux et les arguments du verbe correspondant, autrement dit, à établir un lien entre les arguments d'une nominalisation et la structure argumentale prédicative du verbe de base. Sur le plan du contenu, le projet inclut, d'une part, la prise en compte des principaux arguments du verbe (sujet, complément direct, complément indirect) ainsi que certains compléments verbaux plus secondaires directement liés aux compléments nominaux et, d'autre part, l'élaboration d'une entrée de nominalisation étendue, incluant des informations relatives aux verbes support que souvent accompagnent les nominalisations (ex. *lancer une attaque*, *faire une promenade*).

Le lexique *NomLex* comprend 1 025 entrées lexicales des nominalisations les plus fréquentes issues de différents corpus (entres autres, Brown Corpus, Wall Street Journal). Dans le cadre du projet *NomLex*, il est également prévu d'annoter toutes les nominalisations issues d'un autre corpus, le Penn Treebank, afin d'étendre et de valider les entrées de *NomLex*.

- *NomBank* (Meyers & *al.* 2004)

NomBank[9] est un projet d'annotation sur corpus de l'Université de New York, en lien avec le projet *PropBank*[10] de l'Université de Colorado. L'objectif de *NomBank* est d'analyser les arguments des noms dans le PropBank Corpus, qui est constitué par le Wall Street Journal Corpus du *Penn Treebank*, tout comme *PropBank* vise à y étudier les arguments des verbes. Dans le cadre du processus d'annotation, le projet *Nombank* produit un certain nombre de ressources, dont divers dictionnaires, permettant d'étiqueter les divers arguments et les adjoints des noms candidats, avec l'attribution de rôles en accord avec les parties du discours. Ce projet a commencé en liaison avec le projet *Nomlex* de Catherine MacLeod. Dans cette optique, l'objectif de *Nombank* est de définir et de décrire la structure argumentale des noms de la manière la plus fine et la plus détaillée possible, ce qui implique l'analyse de divers phénomènes tels que les constructions des verbes support, les arguments des copules, les constructions des syntagmes prépositionnels : l'intérêt de cette étude est de constater que l'argument d'un nom peut se trouver en dehors du syntagme nominal dont ce nom

---

[8] Cf. sur *Nomlex* : http://nlp.cs.nyu.edu/nomlex/index.html

[9] Cf. site sur *Nombank* : http://nlp.cs.nyu.edu/meyers/NomBank.html et l'article de Adam MEYERS, Ruth REEVES, Catherine MACLEOD, Rachel SZEKELY, Veronika ZIELINSKA, Brian YOUNG et Ralph GRISHMAN. *The NomBank Project: An Interim Report*. New York University.
Document consultable et téléchargeable en format PDF à l'adresse : http://nlp.cs.nyu.edu/meyers/papers/nombank-pap.pdf

[10] *PropBank*, pour Proposition Bank, est un projet d'annotation sémantique de corpus de texte élaboré par l'Université de Colorado aux Etats-Unis, sous l'égide de Martha Palmer. L'objectif du projet *PropBank* est d'étiqueter les structures argumentales prédicatives dans le *Penn Treebank*, avec des étiquettes relatives au sens et des étiquettes relatives aux arguments qui sont basées sur les entrées lexicales du *Penn Treebank* contenant des informations relatives à leurs structures argumentales prédicatives, les relations entre prédicat et argument ayant été ajoutées aux arbres syntaxiques du *Penn Treebank*.

est la tête. Ce projet vise ainsi à analyser les nominalisations des verbes mais aussi celles des adjectifs. La version 1.0 de *NomBank* est sortie le 17 décembre 2007 : elle couvre tous les noms analysables du Wall Street Journal Corpus du *Penn Treebank*, à savoir, 114 576 propositions et 202 965 occurrences de noms.

- *AnCora* (Taulé & al. , 2008)

Le lexique obtenu dans le cadre du projet AnCora (Taulé & al. , 2008) a été élaboré à partir d'annotations manuelles et semi-automatiques effectuées à différents niveaux. Au niveau sémantique, l'annotation de la structure argumentale verbale a permis l'enrichissement du niveau syntaxique comme l'atteste le tableau ci-dessous récapitulant les fonctions possibles que peut réaliser chaque argument.

Arg0	Agent complement, Direct object, Indirect object, Subject
Arg1	Adjunct, Direct object, Prep. comp., Subject
Arg2	Attribute, Adjunct, Direct object, Indirect object, Predicative, Prep. comp., Subject
Arg3	Adjunct, Indirect object, Predicative
Arg4	Adjunct
ArgM	Adjunct, Predicative
ArgA	Prep. comp., Subject
ArgL	Adjunct, Direct object, Predicative, Prep. comp., Subject

Tableau 6 – fonctions possibles d'un argument dans *AnCora*

L'annotation des rôles thématiques dans le projet AnCora a révélé que chacun de ces arguments pouvait coïncider avec des rôles thématiques spécifiques. Les rôles thématiques qui ont été mis en évidence sont AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State) et ADV (Adverbial). Enfin, concernant l'annotation au niveau sémantique dans le projet AnCora, une annotation manuelle a consisté en l'attribution d'un sens à chaque substantif à partir de WordNet.

4. Evaluation des ressources

- **Protocole**

Choix d'un nom prédicatif pour la prédiction de son contexte droit

Afin d'évaluer les lexiques *Lexique-Grammaire des noms prédicatifs*, *Lefff* et *Nomage*, nous avons choisi de prédire le contexte droit du nom prédicatif *construction*. Le choix de ce nom s'explique par le fait qu'il soit fréquemment utilisé dans la presse et qu'il présente une triple construction. En effet, le nom *construction* peut être employé sans argument (ex : *la construction a pris du retard*), avec un argument introduit par la préposition *de* (ex : *la construction de la maison a pris du retard*) ou deux arguments dont le second est introduit par la préposition *par* (ex : *la construction de la maison par l'entreprise a pris du retard*).

Constitution du corpus de test

Nous avons constitué un corpus de 25 phrases contenant des occurrences de formes fléchies du nom *construction*. Il s'agit de prédire le contexte droit de ces occurrences à l'aide des informations

contenues dans les ressources que nous avons présentées précédemment pour les noms prédicatifs. De la même manière que pour la constitution du corpus de test pour les verbes prédicatifs, nous avons extrait 20 phrases de *l'Est Républicain* (année 2003) illustrant les trois constructions différentes du nom *construction*. Dans ce corpus, nous avons inséré 5 phrases agrammaticales et inacceptables du point de vue du sens.

Corpus :

1. En attendant l'éventuelle construction d'une usine de potabilisation
2. 1.036 euros pour une nouvelle construction.
3. Le toit de la salle thillotine de construction récente avait jusqu'à maintenant résisté à toutes les tempêtes.
4. le projet de construction d'une aire de stockage des boues
5. un éventuel financement pour la construction d'une nouvelle gendarmerie
6. elle participe aux projets de maîtrise d'eau, à la construction de dispensaires.
7. Les travaux de construction seront terminés début 2003.
8. Cela entraînait un surcoût mais avait l'avantage de ne pas allonger les délais de construction
9. la construction par l'état de cinq logements
10. la construction de la plate-forme logistique débutée attend d'éventuels repreneurs
11. l'évocation de la construction prochaine de leur nouvelle caserne
12. Elle prévoit la construction d'un trottoir
13. la construction commencera en 2004
14. la construction de deux classes primaires par la ville d'Héricourt
15. Vous étiez solidaire dans la construction de cette équipe.
16. la construction d'un nouveau centre d'intervention est toujours programmé
17. Quinze places publiques cédées pour la construction des cuisines de l'hôtel
18. En 2004, il y aura la construction de l'école
19. j'envisage la construction d'un atelier
20. Il prévoit la construction dans l'immédiat de deux lotissements
21. *La construction de la plante
22. *La construction manger
23. *La construction de la plante par Max
24. *La construction de la soupe
25. *La construction par Max de l'eau

Données

Nous avons extrait des ressources existantes sur les noms prédicatifs les informations nécessaires à la prédiction du contexte droit du nom *construction*. Pour exemple, nous avons extrait de *Lefff* les informations suivantes :

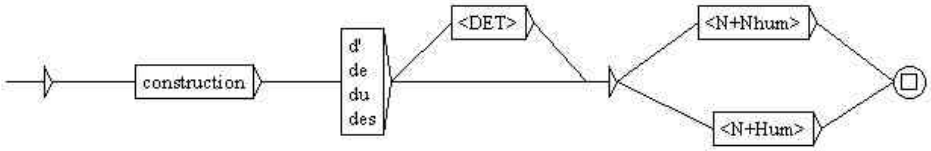
pred="construction_____1<Objde:(de-sinf|de-sn)

Ces informations indiquent que *construction* possède un argument introduit par la préposition *de*. Son argument doit se réaliser syntaxiquement sous la forme d'un syntagme nominal ou d'une proposition infinitive.

Nous avons adaptés par la suite ces informations sous la forme de graphes de reconnaissance de la même manière que pour la prédiction du contexte des verbes prédicatifs et nous avons créé un dictionnaire prenant en compte les informations sémantiques (« +Hum », « +Nhum ») de certaines des ressources.

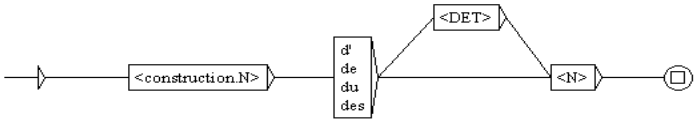
Élaboration des graphes de reconnaissance

– Graphe de *LGLex/Lexique-Grammaire*



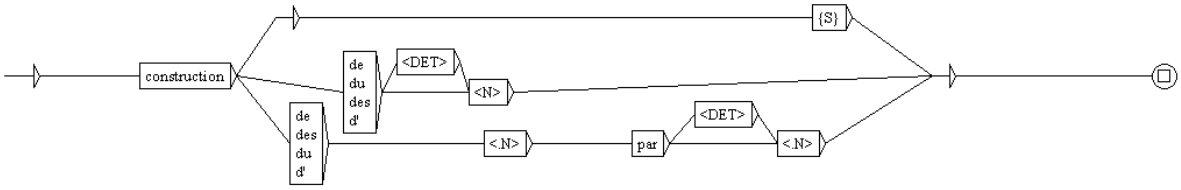
Graphe LGLex - construction

– Graphe de *Lefff*



Graphe Lefff - construction

– Graphe du lexique *Nomage*



Graphe Nomage - construction

- **Résultats**

L'évaluation des ressources que nous venons de présenter pour la prédiction du contexte droit du nom prédicatif *construction* retourne les résultats suivants.

Ressource	Nombre de contextes droits incorrects reconnus (/5)	Nombre de contextes droits reconnus (/25)
Lefff	4	15 (60%)
LGLex/Lexique-Grammaire	4	15 (60%)
Nomage	4	15 (60%)

Tableau 7 - résultats de la prédiction du contexte droit de *construction* à base de connaissances syntaxiques

Ressource	Nombre de contextes droits incorrects reconnus (/5)	Nombre de contextes droits reconnus (/25)
Lefff	4	15 (60%)
LGLex/Lexique-Grammaire	4	15 (60%)
Nomage	4	15 (60%)

Tableau 8 - résultats de la prédiction du contexte droit de *construction* à base de connaissances syntaxiques et sémantiques

Ces résultats sur l'évaluation des ressources existantes pour la prédiction du contexte droit des verbes prédicatifs apportent deux informations intéressantes :

– l'intégration des connaissances sémantiques dans la prédiction du contexte droit du nom prédicatif *construction* n'entraîne pas de modification des taux de prédiction du tableau 8. Cela tient au fait que seule la ressource *LGLex* contient des informations sémantiques sur les arguments du nom *construction* de type traits sémantiques. Par ailleurs, les traits sémantiques en question étant /*+Hum/* (humain) et /*+Nhum/* (non humain), ils s'annulent et permettent la prédiction de tout type de noms sans véritable distinction sémantique.

– Ces trois ressources ont le même taux de prédiction que ce soit à partir de connaissances syntaxiques uniquement ou à partir de connaissances syntaxiques et sémantiques. Ce résultat homogène s'explique par une description identique du cadre de sous-catégorisation de *construction* dans les trois ressources et par le fait qu'aucune de ces ressources ne se distinguent par l'intégration de connaissances sémantiques.

En conclusion, concernant la prédiction du contexte droit des noms prédicatifs, nous ne pouvons pas déterminer tant du point de vue quantitatif et qualitatif quelle ressource permet une meilleure prédiction du fait de l'absence d'une description sémantique des arguments des noms prédicatifs.

Partie 4 : prédiction du contexte droit des adjectifs prédicatifs

1. Définition d'un adjectif prédicatif

Parmi les adjectifs en français, il existe des adjectifs se construisant obligatoirement avec un argument interne comme *désireux, exempt* ou encore *enclin* (Noailly 1999). Les premiers adjectifs, que l'on nomme adjectifs prédicatifs, nous intéressent particulièrement pour la prédiction de leur contexte droit. Mais encore faut-il poser les critères pour les distinguer des autres adjectifs et déterminer la nature de leur contexte droit.

À partir du corpus annoté morpho-syntaxiquement *French Treebank* (Abeillé 2003) constitué d'articles du journal *Le Monde* publiés entre 1989 et 1993, Kupsc (2008) a extrait automatiquement la valence des adjectifs qualificatifs (ou cadre de sous-catégorisation) présents dans des constructions attributives ou prédicatives pouvant potentiellement avoir un ou plusieurs arguments. Cette auteure définit la valence d'un adjectif par le nombre d'arguments nécessaires pour que sa construction soit correcte et la représentation syntaxique de ses arguments.

Pour réaliser cette tâche, un certain nombre de critères sont posés au préalable pour distinguer les cas où l'argument est obligatoire des cas où l'argument de l'adjectif qualificatif est optionnel :

– dans les constructions impersonnelle (*il est agréable que Marie vienne/il est agréable de sortir*), lorsque la clause extrapolée peut être sujet, alors elle est considérée comme l'argument sujet de l'adjectif (*Que Marie vienne est agréable/De sortir est agréable*) ;

– dans d'autres constructions, l'argument de l'adjectif est considéré comme objet lorsqu'il ne peut être déplacé en position sujet (*Paul est heureux que Marie vienne/*Que Marie vienne Paul est heureux*) ;

– dans les constructions comparatives, les adjectifs n'ont pas d'argument obligatoire car les constructions comparatives sont possibles avec presque tous les adjectifs

L'étude de Kupsc (2008) montre que la majorité des adjectifs qualificatifs (86% des adjectifs qualificatifs) ont un cadre de sous-catégorisation dit “simple” possédant un seul argument en fonction sujet (SUJ:NP). Parmi les adjectifs qualificatifs étudiés seuls 61 apparaîssent toujours dans le corpus avec un argument comme *accessoire, accompagné, adhérent, admis, agrégé, allergique, amateur, apte, aride, attenant, avare, concessionnaire, condamné, coupable, coutumier, destructeur, distant, désireux, exempt, fier, fixé, incapable, interdit, semblable.* Finalement, Kupsc (2008) détermine à partir de son expérience que les arguments des adjectifs peuvent se réaliser sous la forme d'un syntagme prépositionnel, d'une proposition subordonnée ou d'une proposition infinitive.

L'extraction automatique de la valence (ou cadre de sous-catégorisation) des adjectifs qualificatifs du *French TreeBank* a permis la constitution de la ressource *TreeLex* que nous présentons dans la section suivante.

D'autres études existent en français sur l'étude des adjectifs prédicatifs mais bien peu face à celles existantes pour les verbes prédicatifs et les noms prédicatifs (voir partie 2 et partie 3). Nous pouvons citer l'étude de Gross 2012 et celle de Léger 2006.

L'étude de Gross (2012) ne propose pas véritablement de classification des adjectifs en fonction de l'étude de leurs arguments mais plutôt une vue d'ensemble des problèmes que pose l'étude des adjectifs prédicatifs. Dans un premier temps, Gross (2012) pose qu'il existe une corrélation entre la signification d'un adjectif et les « schémas d'argument ». Le sens d'un adjectif change en fonction des classes sémantiques de ses arguments. L'adjectif *juste* a la même signification dans *le piano <instrument de musique) est juste* et *le violon <instrument de musique> est juste* mais pas dans *le pantalon <vêtement> est juste, pantalon* étant d'une classe sémantique différente. Gross (2012)

aborde par la suite les propriétés syntaxiques des adjectifs prédicatifs. Il énumère plusieurs structures argumentales rencontrées avec des adjectifs *:* adjectifs sans argument (*cette région est tranquille*), compléments en *à N* (*Paul est fidèle à Marie*), compléments en *de N* (*Paul est conscient du danger*), compléments en *avec N* (*Paul est gentil avec moi*), compléments en *contre N, après N* (*Paul est fâché contre nous, Paul est fâché après nous*), compléments en *pour N* (*ce texte est difficile pour un enfant*), complément en *devant N* (*Paul est admiratif devant cette attitude*), compléments en *en N* (*Paul est fort en maths*), compléments en *(de ce) que P* (*je suis sûr que la réponse est exacte*). Ici, Gross (2012) se contente de relever les constructions possibles et ne propose pas d'étude de leurs propriétés syntaxiques. Enfin, l'auteur liste un certain nombre de problèmes que l'on rencontre pour l'étude de la structure argumentale des adjectifs sans se lancer dans une étude approfondie :

– les arguments des adjectifs sont introduit majoritairement par une préposition sauf parfois par des complétives directes en *que P ;*

– les classes d'objet permettent de prédire la préposition qui introduit l'argument : *paul est bon au bridge <jeu>*. Si l'argument appartient à la classe sémantique <jeu> alors on peut prédire que la préposition introductive est *à ;*

– l'effacement de l'argument de l'adjectif est possible mais peut provoquer un changement de signification de l'adjectif : *la maison est haute de 3 mètres/maison est haute* (évaluation subjective de l'individu) ;

– il existe un problème de polarité au sein des couples antonymiques : c*ette planche est longue de 3 mètres/*cette planche est courte de 3 mètres.*

L'étude de Léger 2006 propose une classification sémantique des adjectifs reflétant également leurs propriétés syntaxiques. Cette classification repose essentiellement sur le critère du mode et met en évidence trois classes d'adjectifs : ceux qui acceptent des propositions infinitives ou des propositions dans lesquelles le verbe est à l'indicatif *(Jean est certain que Marie partira/Jean est certain de partir*), ceux qui acceptent des propositions infinitives ou des propositions dans lesquelles le verbe est au subjonctif (*Jean est content que Marie parte/Jean est content de venir*) et ceux qui acceptent uniquement des propositions infinitive (*Jean est facile à convaincre/*Jean est facile qu'on convainc*). Les résultats obtenus par Kupsc (2008) semble se conformer à la classification de Léger 2006.

2. Ressources existantes

▪ *TreeLex*

L'extraction automatique des cadres de sous-catégorisation des adjectifs qualificatifs au sein du corpus annoté *French Treebank* et la définition des critères permettant de distinguer les arguments obligatoires des arguments optionnels ont permis de constituer le lexique *TreeLex.* Ce lexique contient des informations sur les cadres de sous-catégorisation des adjectifs prédicatifs telles que :

- **SUJ** désigne la fonction sujet d'un argument (realisée comme un syntagme nominal, une proposition infinitive ou une proposition subordonnée)
- **OBJ** désigne la fonction objet d'un argument (réalisée comme une proposition subordonnée)
- **P-OBJ** désigne la fonction objet indirect (réalisée comme syntagme prépositionnel ou proposition infinitive introduite par une préposition)
- **cl** désigne un clitique

Voici un extrait de *TreeLex* pour illustrer ces différentes fonctions :
nécessaire
|SUJ:NP (38)

|SUJ:NP|P-OBJ:PP[a] (14)
|SUJ:NP|P-OBJ:VPinf[pour] (9)
|SUJ:VPinf[de] (3)
|SUJ:NP|P-OBJ:VPinf[de] (2)
|SUJ:NP|P-OBJ:PP[pour] (2)

Le lexique *TreeLex* regroupe actuellement 271 adjectifs accompagnés de 27 cadres de sous-catégorisation différents.

- ***Lexique-Grammaire/LGLex/Lefff***

La présence d'adjectifs dans la ressource *Lefff,* que nous avons d'ailleurs déjà présentée dans la partie 2 de ce dossier, provient de la fusion avec la ressource *LGLex* qui est elle-même l'adaptation notamment des tables des adjectifs prédicatifs du *Lexique-Grammaire.*

3. Evaluation des ressources

- **Protocole**

Concernant l'évaluation des ressources *Lefff* et *TreeLex*, nous suivons le même protocole utilisé pour l'évaluation des ressources sur les verbes et les noms prédicatifs.

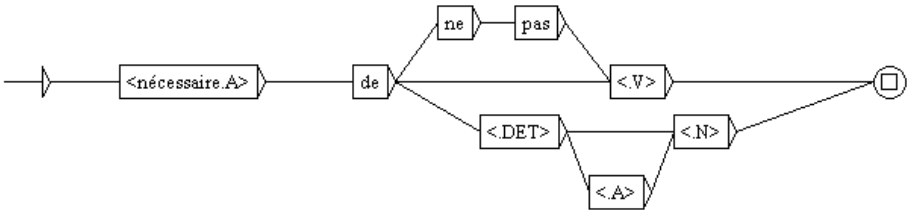
Il s'agit ici de prédire le contexte droit de l'adjectif prédicatif *nécessaire*. Cet adjectif est d'usage fréquent et possède 6 constructions différentes (*cf.* informations extraites de *TreeLex* ci-dessus). Cet adjectif peut s'employer (i) sans argument, avec un argument de type syntagme nominal introduit par la préposition (ii) *à* ou (iii) *de,* avec un argument de type proposition infinitive introduit par la préposition (iv) *à* ou (v) *de* ou avec un argument de type syntagme nominal introduit par la préposition (vi) *pour.*

Corpus :

1. l'utilisation d'un tel matériel s'avérera nécessaire.
2. Les 400.000 plants nécessaires au fleurissement d'été
3. il est nécessaire de réserver son repas avant le 10 janvier.
4. Un trempage à l'eau tiède suivi d'une bonne douche se sont révélés nécessaires pour rendre une apparence normale
5. Toutes les signalisations nécessaires seront mises en place par les services municipaux.
6. Bruno Warnet puisera calme et sérénité nécessaires à son équilibre
7. Une mutualisation nécessaire de la ressource
8. Ce travail est nécessaire pour améliorer notre vie quotidienne.
9. Une bonne demi-heure et plusieurs allers-retours sont nécessaires à la passagère
10. Une journée de familiarisation est nécessaire
11. Trois cuissons sont nécessaires pour appliquer de l'or
12. Nous demandons la poursuite des travaux d'aménagements nécessaires au fonctionnement
13. Plus de 3,75 millions de voix nécessaires
14. Il est salutaire et nécessaire de ne pas baisser les bras
15. le Conseil général fonde de grands espoirs et entend consacrer les moyens nécessaires à leur réalisation
16. les ressources nécessaires à la réalisation des projets
17. Tous les services nécessaires au bon déroulement du mariage sont présents
18. Autour des spectacles, une mise en scène plastique est nécessaire.
19. l'alternance des entraînements et des compétitions s'avère nécessaire pour l'obtention des grades
20. Afin que ces 6 millions d'euros d'investissement et le suivi effectué par le préfet ne soient pas seulement nécessaires à Neuves-Maisons
21. *Max a acheté un agenda nécessaire pour fermer la porte
22. *Max demande une aide financière nécessaire à la peur
23. *Max demande une aide financière nécessaire pour sa phrase
24. *Max a acheté un agenda nécessaire indispensable
25. *Max demande une aide financière nécessaire fermer

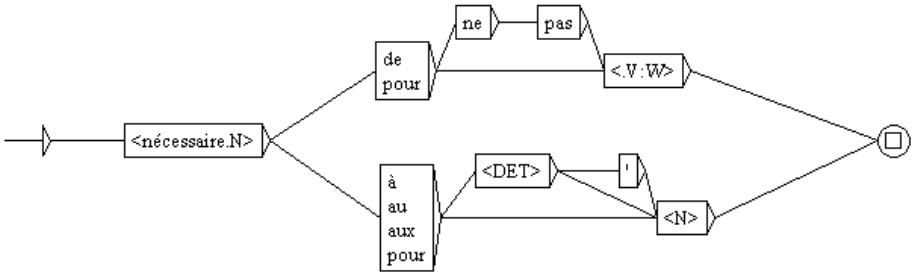
Élaboration des graphes de reconnaissance :

- **Graphe pour *Lefff***



Graphe Lefff - nécessaire

- **Graphe pour *TreeLex***



Graphe TreeLex - nécessaire

- **Résultats**

L'évaluation des ressources que nous venons de présenter pour la prédiction du contexte droit de l'adjectif prédicatif *nécessaire* retourne les résultats suivants.

Ressource	**Nombre de contextes droits incorrects reconnus (/5)**	**Nombre de contextes droits reconnus (/25)**
TreeLex	3	16 (64%)
Lefff	0	3 (12%)

Tableau 9 - résultats de la prédiction du contexte droit de *nécessaire* à base de connaissances syntaxiques

Ces résultats sur l'évaluation des ressources existantes pour la prédiction du contexte droit des adjectifs prédicatifs apportent montre que d'un point de vue purement quantitatif, la ressource qui offre le meilleur taux de prédiction du contexte droit de l'adjectif prédicatif *nécessaire* est *TreeLex* avec un taux de 61%. En revanche, la ressource *Lefff* a un taux de prédiction moins élevé avec 12%. Cela s'explique par une description des cadres de sous-catégorisation possibles pour l'adjectif

nécessaire plus complète dans *TreeLex* que dans *Lefff*. Dans la ressource *TreeLex*, l'adjectif est décrit selon 5 cadres de sous-catégorisation différents alors que *Lefff* n'en décrit que 2.

Partie 5 : Ressources commerciales disponibles au catalogue de l'ELDA

Dans cette partie, nous fournissons une évaluation qualitative des ressources commerciales disponibles au catalogue de l'ELDA (http://www.elda.org/), pertinentes au regard des objectifs du projet Sémoteur. Les différentes ressources sont désignées par leur intitulé au catalogue de l'ELDA, les éléments de présentation synthétique sont en partie ceux fournis par le catalogue, ou par les éditeurs de ressource. Les ressources examinées ci-après ont été sélectionnées sur les critères suivants :

- lexique électronique en langue française ;
- informations syntaxiques sur les entrées verbales, nominales ou adjectivales, indiquant le nombre (valence) et le type de compléments (cadre de sous-catégorisation) ;
- informations sémantiques sur les entrées présentes dans la ressource.

Les tarifs indiqués sont ceux relevés dans le catalogue ELDA.

ELRA-L0005 Lexique Français

Les informations ci-dessous sont tirées du catalogue ELDA;

- Vocabulaire général
- Entrées : 50 000
- Format : ASCII
- Support : disquette, cartouche QIC 150 MB

Le lexique français lanTmark se répartit selon les catégories suivantes : noms (36.000), verbes (6.000), adjectifs (7.000), adverbes (1.000).
Chaque entrée comporte des informations morphologiques (flexions, marques de superlatif et comparatif), informations syntaxiques (traits de position, genre, marqueurs de complément et arguments de verbe), informations sémantiques (lexico-sémantiques pour les noms et les adjectifs).

Prix Membres	**Prix Non Membres**
Academic - Commercial 48000.00 EUR	Academic - Commercial 80000.00 EUR
Academic - Research 6000.00 EUR	Academic - Research 10000.00 EUR
Commercial - Commercial 48000.00 EUR	Commercial - Commercial 80000.00 EUR
Commercial - Research 48000.00 EUR	Commercial - Research 80000.00 EUR

Évaluation

L'intérêt de cette ressource réside dans la richesse d'informations disponibles, notamment aux niveaux syntaxique et sémantique : « marqueurs de complément et arguments de verbe » et informations « lexico-sémantiques pour les noms et les adjectifs ». Toutefois, aucun échantillon ou exemple de description ne sont fournis. Par ailleurs, la date d'édition (1997), ainsi que les supports de stockage indiqués laissent augurer d'une ressource ancienne, qui n'a pas fait l'objet de mises à jour récentes. Le prix demandé pour cette ressource, au regard de son caractère relativement obsolète, semble exagéré.

ELRA-M0001 Lexique multilingue de base (MEMODATA)

Les informations ci-dessous sont tirées du catalogue ELDA.

- Entrées : 30 000 pour chaque langue
- Langues : français, anglais, italien, allemand, espagnol
- Format: ASCII ou ANSI avec séparateurs entre les entrées
- Support : CD-ROM

Les mots sont associés par leur sens. Les catégories lexicales sont : noms (5 * 18 000), verbes (5 * 8 000), adjectifs (5 * 6 000), adverbes (5 * 1 500).

Des échantillons de la ressource sont disponibles dans le catalogue ELDA pour les verbes commençant par « ca- » : http://www.elda.org/catalogue/fr/text/doc/lexmult.html.

- Échantillon d'informations multilingues pour « calmer »

Entrée FR	gr. FR	DE	gr. DE	EN	gr. EN	IT	gr. IT	ES	gr. ES
cacher	v_trans	verbergen	v_trans	hide	v_trans	nascondere	v_trans	disimular	v_trans
cacher	v_trans	verbergen	v_trans	hide	v_trans	nascondere	v_trans	esconder	v_trans
cacher	v_trans	verhehlen	v_trans	conceal	v_trans	nascondere	v_trans	ocultar	v_trans
cacher	v_trans	wegstecken	v_trans	hide away	v_trans	nascondere	v_trans	esconder	v_trans

Prix Membres	Prix Non Membres
Academic - Commercial 11077.00 EUR	Academic - Commercial 13846.00 EUR
Academic - Research 8861.00 EUR	Academic - Research 11077.00 EUR
Commercial - Commercial 11077.00 EUR	Commercial - Commercial 13846.00 EUR
Commercial - Research 11077.00 EUR	Commercial - Research 13846.00 EUR

Comme on peut le voir dans le tableau ci-dessus, chaque entrée pointe vers les 4 autres langues proposées. Les entrées sont par ailleurs associées à un code décrivant de façon succincte leur cadre de sous-catégorisation : « v_trans », pour verbe transitif, soit un verbe attendant un complément direct. Toutefois, aucune précision n'est donnée sur la réalisation syntaxique du complément.
En complément de ces éléments, le site http://www.memodata.com fournit des démonstrateurs pour les différentes ressources et logiciels développés par l'entreprise.

- Exemple d'informations disponibles pour le verbe « manger »

manger (v.)
eat, feed

manger (v.) (V+comp)
eat, eat up
voir aussi

manger (v.)

↘ becquetage, bouffable, mangeable, mangeur
Le Dictionnaire Intégral & Wordnet

manger (n. m.)
nourriture (pour l'homme)[Classe]

manger (v. intr.)
accomplir une action[Classe]

manger — eat up; eat[Classe]
consommer, ingérer, prendre — consume, have, ingest, take, take in[Hyper.]

alimentation — eating, feeding - mangeur — eater, feeder[Dérivé]
becqueter, becter, béqueter, boulotter, croûter, gnafrer, grailler, manger, tortorer — eat, eat up[Domaine]

manger (v. tr.)
manger — eat up; eat[ClasseHyper.]
consommer, ingérer, prendre — consume, have, ingest, take, take in - manger — eat[Hyper.]

alimentation — eating, feeding - comestibles — eater - mangeur — feeder[Dérivé]
bouffer, nourrir — feed - mâcher, mãcher, mastiquer — chew, chew on, jaw, manducate, masticate - absorber, avaler — get down, swallow, swallow down[Domaine]

manger (v. tr.) [figuré]
dépenser[Classe]

abîmer par usure[Classe]

manger (verbe)
consommer, ingérer, prendre — consume, have, ingest, take, take in[Hyper.]

alimentation — eating, feeding - aliment, biberon, fourrage, nourriture — feed, provender - mangeur — eater, feeder[Dérivé]

becqueter, becter, béqueter, boulotter, croûter, gnafrer, grailler, manger, tortorer — eat, eat up[Domaine]
locutions

See conjugation of manger • aliments prêts à manger • baguettes à manger chinoises • besoin de manger • bien manger • bien-manger • blanc-manger • boire et manger • bête à manger du foin • coin salle à manger • donner à manger • entre-manger • faire manger • faire à manger • finir (de manger) • garde manger • garde-manger • le boire et le manger • manger au râtelier • manger avec un lance-pierres • manger comme quatre • manger comme un chancre • manger comme un moineau • manger de baisers • manger de bon appétit • manger de la vache enragée • manger des bonbons • manger des yeux • manger doucement • manger du bout des dents • manger en faisant du bruit • manger en paissant • manger la grenouille • manger le morceau • manger le sang • manger les pissenlits par la racine • manger ses mots • manger ses quatre sous • manger son blé en herbe • manger son pain blanc • manger son pain noir • manger son patrimoine • manger sur le pouce • manger tout son soûl • manger un morceau • manger à belles dents • manger à la fortune du pot •

manger à tous les râteliers • mobilier de salle à manger • perdre le boire et le manger • préposé à la table à manger • prêt à manger • prêt-à-manger • salle (à manger) • salle à manger • salle à manger table (e) • salon-salle à manger • table de salle à manger

BLANC-MANGER • ENTRE-MANGER • GARDE-MANGER[Littré]

Blanc-manger • Comment manger avec son cul • M.I.A.M : Mon Invitation À Manger • Manger bouger • Pret A Manger • Roger Ebert devrait manger moins gras • Salle à manger • Salle à manger d'État (Maison Blanche)

L'intérêt principal de la ressource Memodata est la complétude des informations lexicales, intégrant une partie des informations indispensables au projet Sémoteur sur la structure argumentale : comme dans l'extrait vu plus haut, le dictionnaire Alexandria dont sont tirées les informations associées à l'entrée « manger » n'indique que le type général de cadre de sous-catégorisation du verbe, sans préciser la nature syntaxique exacte des compléments attendus. L'autre intérêt de ces ressources est la possibilité de naviguer au sein d'un réseau sémantique dense, et de passer ainsi d'une entrée verbale « manger » à par ex. des hyperonymes : « ingérer, consommer » tant en français que dans les autres langues disponibles. D'autres relations sémantiques sont exploitables, ce qui accroît encore la couverture de la ressource. Ainsi, une distinction est faite entre acceptions littérales et figurées, mais d'autres locutions intégrant « manger » ou liées au concept lui-même sont également disponibles. La ressource Memodata donne également accès à des entrées relevant d'autres catégories syntaxiques mais participant du réseau sémantique : des noms génériques tels que « nourriture » ou encore des noms spécifiques tels que « becquetage », des adjectifs tels que, « mangeable, mangeur » ou encore « bouffable ». En complément des éléments morphologiques, syntaxiques et sémantiques, la ressource donne ainsi accès à des niveaux de langue différents, ce dans plusieurs langues.

Évaluation

L'entreprise Memodata produit un certain nombre de ressources dictionnairiques utilisables, dont leur dictionnaire de base « Dictionnaire intégral », se déclinant en une API appelée « Sémiographe ». Cette API, basée sur une compilation du Dictionnaire Intégral, fournit un très grand nombre de services améliorant le développement d'applications linguistiques. Le Sémiographe contient les langues d'Alexandria, le démonstrateur disponible sur http://www.memodata.com et dont nous avons extrait les informations ci-dessus pour « manger », et près de cinquante API (de gestion de mots (lemmatisation, filtrage de sens), de phrases (expansion de requête, POS tagger), et de textes (résumé, routage etc.). De ce fait, la ressource proposée par Memodata présente de nombreuses caractéristiques compatibles avec les objectifs du projet Sémoteur. Toutefois, l'API est annoncée comme étant « [e]n refonte en ce moment, ce produit n'est plus disponible à la vente pour quelques mois. L'ancienne version peut néanmoins être obtenue. » Cet avis est affiché depuis plusieurs années, d'après nos informations.

Le prix affiché pour l'API du Sémiographe est : « à partir de 2900 €HT », avec les réserves exprimées précédemment sur l'ancienneté de l'information. Signalons que les tentatives de prise de contact afin de déterminer l'état de la ressource et ses modes de commercialisation sont restés, à ce jour, sans réponse.

ELRA-M0033 SCI-FRAN-EURADIC Dictionnaire bilingue français-anglais

Ce dictionnaire bilingue a été augmenté et amélioré dans le cadre du projet national français EuRADic (Dictionnaire et corpus européens et arabe), du programme Technolangue, financé par le

ministère français de l'industrie. Il contient environ 243 539 couples de termes français-anglais, avec leur partie du discours. Les données sont présentées sous un format tabulaire et les informations relatives à chaque entrée séparées par des ";".
D'autres formats et services peuvent être proposés par le fournisseur de ces données à la demande (par exemple, conversion vers le formalisme de l'acquéreur, sélection de sous-ensembles de mots en complément de dictionnaires existants).

Une description du projet est disponible à l'adresse suivante : http://www.technolangue.net/article.php3?id_article=203.

Voir aussi : ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0034, ELRA-M0035, ELRA-M0036, ELRA-M0037, ELRA-M0038.

Prix Membres	Prix Non Membres
Academic - Commercial 8000.00 EUR	Academic - Commercial 11000.00 EUR
Academic - Research 1000.00 EUR	Academic - Research 1800.00 EUR
Commercial - Commercial 8000.00 EUR	Commercial - Commercial 11000.00 EUR
Commercial - Research 8000.00 EUR	Commercial - Research 11000.00 EUR

Prix Spéciaux :
Réductions offertes pour l'achat de plusieurs ressources SCIPER (L0049, L0050, L0051, L0052, L0053, M0033, M0034, M0035, M0036 et M0037):
• 10% de réduction sur l'achat de 2 dictionnaires,
• 20% de réduction sur l'achat de 3 dictionnaires,
• 25% de réduction pour l'achat de plus de 3 dictionnaires.

Évaluation

D'après les informations disponibles, l'intérêt principal de cette ressource réside dans son caractère bilingue, et dans l'indication de leur catégorie syntaxique. Cette ressource peut donc être rapidement intégrée dans une perspective de réaliser des transferts lexicaux (traduction automatique mot-à-mot), toutefois l'absence d'informations sur les cadres de sous-catégorisation des unités lexicales décrites supposerait le croisement de cette ressource avec d'autres ressources plus complètes, telles que Dicovalence, Treelex ou encore Nomage, pour le français, au minimum Nombank et Wordnet pour l'anglais. En l'état, cette ressource ne présente qu'un intérêt relatif, étant donné les objectifs du projet Sémoteur. Toutefois, elle peut être envisagée comme une ressource intéressante dans l'optique d'un développement multilingue des applications de la SAS Ergonotics. Cependant, les tarifs affichés pour la ressource apparaissent relativement excessifs au regard des informations lexicales disponibles.

ELRA-M0020 EuroWordNet français

Le projet Wordnet est un réseau sémantique constitué selon des principes psycholinguistiques. Le premier wordnet a été élaboré par l'université de Princeton dans les années 1990, sous régime de licence gratuit et, dans ses dernières versions, ouvert. Contrairement au wordnet de Princeton, EurowordNet est le fruit d'un programme de recherche et développement impliquant un consortium d'entreprises et de laboratoires universitaires européens, et est diffusé uniquement sous licence commerciale. Les applications du réseau sémantique WordNet1.5 de Princeton sont multiples en

Traitement Automatique des Langues et Recherche d'Information, étant donné qu'aux informations proprement lexicographiques (parties du discours, propriétés morphologiques, cadre de sous-catégorisation) sont associées des informations sémantiques (relations sémantiques : synonymie, antonymie, hyper- et hyponymie au minimum), ainsi que des gloses (semi-définitions en langage naturel) et des exemples. La version française de EurowordNet est alignée sur les versions réalisées dans les autres langues européennes, ce qui permet de disposer, en complément d'une ressource lexicale riche, d'une ressource permettant d'envisager des traitements multilingues. Toutefois, comme indiqué ci-après, les entrées de EurowordNet ne comportent pas d'informations sur la sous-catégorisation des éléments prédicatifs (Verbes, Adjectifs et Noms).

Les informations ci-dessous sont extraites du catalogue de l'ELDA.

- Wordnets disponibles :

ELRA ref.	Langue	Synsets	Sens des mots	Relations internes à la langue	Relations d'équivalence
ELRA-M0015	Anglais: Addition au WordNet anglais	16361	40588	42140	0
ELRA-M0016	Hollandais	44015	70201	111639	53448
ELRA-M0017	Espagnol	23370	50526	55163	21236
ELRA-M0018	Italien	48529	48499	117068	71789
ELRA-M0019	Allemand	15132	20453	34818	16347
ELRA-M0020	Français	22745	32809	49494	22730
ELRA-M0021	Tchèque	12824	19949	26259	12824
ELRA-M0022	Estonien	9317	13839	16318	9004

- Composants communs

Une sélection d'enregistrements de l'index inter-lingue, concepts de base, qui jouent un rôle majeur dans les différents wordnets. Ces concepts de base forment le noyau de tous les wordnets. Tous les concepts de base sont classés en termes de concepts supérieurs qui s'y appliquent.

L'index inter-lingue, qui consiste en une liste d'enregistrements sous la forme de "synsets" (ensembles/réseaux sémantiques, principalement issus de WordNet5.1 ou créés manuellement), comprend :

- A.1. un ensemble de synsets de mots ou phrases synonymiques (provenant pour la plupart de WordNet1.5) ;
- A.2. une "partie-du-discours" ;
- A.3. un ou plusieurs concepts supérieurs (optionnel) ;
- A.4. un ou plusieurs étiquettes de domaine (optionnel) ;
- A.5. un glossaire en anglais (provenant pour la plupart de WordNet1.5) ;
- A.6. un code unique reliant le synset à sa source (provenant pour la plupart de WordNet1.5).
- Ontologie supérieure : une ontologie de 63 classes sémantiques de base reposant sur des distinctions fondamentales. Grâce à l'ontologie supérieure, on accède à tous les wordnets en utilisant un schéma de classifiaction unique indépendant de la langue. Les concepts supérieurs sont également assignés aux enregistrements de l'index inter-lingue.
- Ontologie de domaine : une ontologie de domaines sujets assignés aux enregistrements de l'index inter-lingue

- Une sélection d'enregistrements de l'index inter-lingue, concepts de base, qui jouent une rôle majeur dans les différents wordnets. Ces concepts de base forment le noyau de tous les wordnets. Tous les concepts de base sont classés en termes de concepts supérieurs qui s'y appliquent.
- WordNet1.5 (91591 synsets; 168217 sens; 126520 mots d'entrée) au format EuroWordNet.

Toutes les données sont distribuées en fichiers textes dans le format EuroWordNet et sous la forme de fichiers de base de données Polaris (voir LR3 ci-dessous). Le visualiseur EuroWordNet (Periscope, voir LR3) peut être utilisé pour accéder à la version base de données. Pour modifier et étendre la version de la base de données, il faut acquérir une licence Polaris.

Les wordnets ne contiennent pas de :

- glossaires
- étiquettes d'usage
- propriétés morpho-syntaxiques
- exemples
- traductions mot-à-mot
- LR(3) Logiciels

La base de données multilingue EUROWORDNET est composée de trois parties :

Les wordnets au format base de données Flaim : un format Novell d'indexation et de compression.

- Polaris (Louw 1997): un éditeur pour la création, l'édition et l'exportation de wordnets.

- Periscope (Cuypers and Adriaens 1997) : un outil graphique pour la visualisation et l'exportation de wordnets.

Polaris peut importer de nouveaux wordnets ou des fragments de wordnets depuis des fichiers ASCII avec le format d'importation correct et crée une base de données indexée EUROWORDNET. De plus, il permet à un utilisateur d'éditer et d'ajouter des relations dans les wordnets et de formuler des requêtes. Polaris rend possible la visualisation de relations sémantiques sous la forme d'une structure arborescente qui peut être directement éditée. Ces arborescences peuvent être étendues et raccourcies en cliquant sur les sens du mot et en spécifiant des "TABs" qui indiquent le type et la profondeur des relations qui doivent être montrées. Les arbres étendus ou les sous-arbres peuvent être stockés sous un ensemble de synsets, qui peuvent être maniés, sauvegardés ou chargés. Il est également possible d'accéder à l'index inter-lingue ou aux ontologies, et de passer des wordnets aux ontologies via l'index inter-lingue. Enfin, il contient une interface permettant de projeter les ensembles de synsets à travers les wordnets.

Le logiciel Periscope est un visualiseur public qui peut être utilisé pour regarder les wordnets créés par Polaris et pour les comparer dans une interface graphique. Les sens des mots peuvent être visualisés et les arborescences étendues. Les sens individuels ou des branches entières peuvent être projetées sur un autre wordnet ou des structure de wordnets peuvent être comparées via les relations d'équivalence avec l'index inter-lingue. Les arbres sélectionnés peuvent être exportés vers des fichiers textes. Periscope ne peut pas importer ou changer les wordnets.

Le programme Polaris est la propriété de Vantage Research et est mis à disposition en tant que résultat d'EuroWordNet à Vantage Research (www.vantage.com).

Le logiciel Periscope est la propriété de Vantage Research.

- Prix

Les prix sont basés sur le nombre de synsets pour chaque langue. Les membres bénéficient d'une remise de 50% sur le prix public. Chaque langue comprend un nombre fixe et indivisible de synsets.

Il y a 4 types différents d'usage :

VAR-C = Usage commercial

VAR-I = Usage interne pour une organisation commerciale

VAR-E = Licence d'évalutation (licence limitée à une durée de 3 mois)

End-User = Usage de recherche par une institution académique

Langue	Nombre de synsets
ELRA-M0015 Anglais (complément)	16 361
ELRA-M0016 Hollandais	44 015
ELRA-M0017 Espagnol	23 370
ELRA-M0018 Italien	48 529
ELRA-M0019 Allemand	15 132
ELRA-M0020 Français	22 745
ELRA-M0021 Tchèque	12 824
ELRA-M0022 Estonien	93172

Remise***

Nombre de synsets	Remise
Au-delà de 60 000 synsets cumulés	5 %
Au-delà de 100 000 synsets cumulés	10 %
Au-delà de 160 000 synsets cumulés	20 %

***Une remise est offerte à la fois aux membres et aux non membres selon le nombre total (cumulé) de synsets faisant l'objet d'une même commande.

Le nombre total de synsets est calculé en additionnant le nombre de synsets de chaque langue achetée. Par exemple, si vous commandez les wordnets anglais et hollandais, le montant total de synsets sera 16 361 synsets (anglais) + 44 015 synsets (hollandais) = 60 376 synsets. Dans ce cas, la remise correspondante de 5 % sera appliquée.

Prix Membres	Prix Non Membres
Academic - Commercial 5686.25 EUR	Academic - Commercial 11372.50 EUR
Academic - Research 227.45 EUR	Academic - Research 454.90 EUR
Commercial - Commercial 5686.25 EUR	Commercial - Commercial 11372.50 EUR

Commercial - Evaluation 454.90 EUR	Commercial - Evaluation 909.80 EUR
Commercial - Research 3411.75 EUR	Commercial - Research 6823.50 EUR

La ressource EurowordNet pourrait correspondre en partie aux objectifs du projet Sémoteur, notamment par son aspect « réseau sémantique » . Toutefois, les entrées lexicales étant dépourvues d'informations sur les cadres de sous-catégorisation des éléments prédicatifs, contrairement à WordNet1.5 de Princeton, en complément de cette ressource il serait nécessaire d'intégrer des informations tirées des ressources universitaires citées par ailleurs : Dicovalence, Treelex, et tables du lexique-grammaire pour les verbes français, Nomage et Treelex pour les noms et les adjectifs.

Synthèse-conclusion : ressources commerciales

Il ressort de l'examen des ressources commerciales disponibles au catalogue de l'ELDA, le seul fournisseur commercial de ressources lexicales électroniques, qu'aucune ressource ne couvre complètement les besoins de la SAS Ergonotics. En effet, aucune des ressources commerciales examinées ne décrit spécifiquement les propriétés de sous-catégorisation des éléments prédicatifs (Verbes, Adjectifs et Noms), à part Memodata. Malheureusement, l'état de cette dernière ressource est indéterminé : les tentatives de prise de contact sont restées sans réponse. De ce fait, notre recommandation, dans le cadre du contrat de prestation de service unissant l'UMR STL à Ergonotics, est de poursuivre l'effort de recherche & développement consenti jusqu'à présent par la SAS Ergonotics, en prenant contact individuellement avec chaque auteur des ressources universitaires examinées :

- Dicovalence
- Lefff
- Lexique des Verbes du Français
- LexSchem
- LexValf
- LGLex
- SynLex
- Tables du lexique-grammaire
- Treelex.

Comme nous l'avons vu, seules les ressources académiques décrivent de façon précise les propriétés de sous-catégorisation des différents éléments prédicatifs, dont nous avons vu le rôle crucial pour la prédiction du contexte droit d'un énoncé, dans le cadre d'une application d'aide à la saisie, et plus largement d'aide à la communication. Il reste donc un effort important pour :

- recenser les unités prédicatives décrites dans chaque ressource ;
- mettre en cohérence les descriptions des propriétés de sous-catégorisation proposées par chaque ressource ;
- établir des liens entre verbes, adjectifs et noms décrits, sur la base de propriétés morphologiques, ex : calme (Adj) Ø ← calme (N) [Ø, SUJ{SN, PRO}]→ calmer (V)

[SUJ{SN, PRO, Comp, Vinf}, OBJ{SN, PRO} ; ici le nom « calme » peut se construire seul ou avec un complément de nom dans lequel le SN a une fonction de sujet pour la phrase (*le calme de Pierre*), il peut être associé à l'adjectif « calme » qui n'attend pas de complément (contrairement à « fier », dans *Max est fier de Luc*), il peut également être associé au verbe « calmer », qui se construit en suivant un schéma classique pour les verbes transitifs, c'est-à-dire un un syntagme nominal sujet, une complétive (*que sa maman le berce calme le bébé*) ou encore un verbe à l'infinitif (*crier le calme*). L'objet peut être réalisé par un SN ou un pronom (*ça le calme*) ;

- identifier les lacunes dans la couverture des différentes ressources ;
- élaborer une ressource lexicale électronique unifiée, décrivant les propriétés de sous-catégorisation des verbes, noms et adjectifs en français.

Nous évaluons à 18 hommes/mois l'effort nécessaire pour réaliser le programme ci-dessus pour un échantillon limité aux 3 000 verbes, adjectifs et noms les plus fréquents du français (soit 9 000 entrées au total). Cette évaluation est proposée sous réserve de pouvoir exploiter les ressources universitaires examinées dans le cadre de cette étude. Un délai supplémentaire de concertation avec les ayant-droits et leurs tutelles de rattachement sera nécessaires le cas échéant : certaines ressources élaborées dans un cadre universitaire ont un statut indéterminé en ce qui concerne leurs possibilités d'exploitation commerciale. Ci-dessous nous associons ressources lexicales et régime de licence, sur la foi des informations disponibles sur les sites web de leurs auteurs :

- licence LGPL-LR (Lesser General Public License For Linguistic Resources) : Dicovalence, Lefff, LexSchem, LGLex, SynLex, Lexique des Verbes du Français ;
- autre (non spécifié) : LexValf, Tables du lexique-grammaire, Treelex.

Conclusion générale

L'entreprise Ergonotics souhaite améliorer l'interaction en langage naturel de ses applications iPhone/iPad par l'implémentation d'un moteur de prédiction du contexte droit des mots à base de connaissances linguistiques. Dans ce contexte, l'objectif scientifique du projet *SéMoteur* était de lancer une réflexion sur le(s) type(s) d'information nécessaire(s) à la prédiction du contexte droit et d'évaluer les ressources linguistiques existantes contenant les types d'information identifiés. Notre réflexion s'est portée spécifiquement sur la prédiction du contexte droit des catégories prédicatives en français, autrement dit les mots acceptant dans leur construction au moins un argument (complément) comme *commencer* pour les verbes prédicatifs, *construction* pour les noms prédicatifs et *nécessaire* pour les adjectifs prédicatifs.

Les types d'information identifiés pouvant être utilisés dans la prédiction du contexte droit des verbes prédicatifs sont la valence, le type de construction (transitive, intransitive, transitive directe, transitive indirecte, double construction), le cadre de sous-catégorisation et la structure argumentale. Ces notions sont étroitement liées à la notion de prédicat verbal mais nous avons vu dans cette étude qu'elles s'appliquent également pour la description des noms et adjectifs prédicatifs. Toutefois, comme l'atteste le nombre de ressources existantes, nous avons une connaissance moins aboutie des mécanismes de restriction des arguments pour les noms et les adjectifs prédicatifs.

Concernant l'évaluation des ressources existantes en français contenant ces types d'information pour les verbes, les noms et les adjectifs prédicatifs, nous avons mis en évidence que (i) l'intégration d'une première couche de connaissances syntaxiques sur les arguments suffit à elle-seule à améliorer le taux de prédiction par rapport à la méthode *n-grammes*, (ii) l'intégration d'une seconde couche de connaissances sémantiques permet de restreindre le nombre d'arguments possibles mais (iii) ces connaissances sont insuffisamment prises en compte dans les ressources existantes. Nous avons pu montrer dans cette étude qu'aucune ressource commerciale ne couvre complètement les besoins exprimées par la SAS Ergonotics, ce qui nous amène à recommander l'intégration de ressources lexicales électroniques élaborées dans un cadre universitaire, en procédant à une fusion des informations lexicales existantes. L'évaluation des ressources universitaires pour la prédiction du contexte droit de *commencer*, *construction* et *nécessaire* montre que la ressource la plus adéquate pour la prédiction du contexte droit des verbes prédicatifs est *LGLex*, pour la prédiction du contexte droit des adjectifs prédicatifs est *TreeLex*. En revanche, nous n'avons pu déterminer quelle ressource est la plus adéquate pour la prédiction du contexte droit des noms prédicatifs, les ressources évaluées présentant le même taux de prédiction.

Cette étude de la prédiction du contexte droit des catégories prédicatives offre plusieurs perspectives de recherche :

(i) la constitution d'une ressource pour les noms et les adjectifs prédicatifs contenant des informations sémantiques sur les arguments ;

(ii) les classes sémantiques de Gross (2012) sont une piste à explorer concernant le typage des arguments des catégories prédicatives dans la mesure où *LGLex* est la ressource qui permet une meilleure prédiction du contexte droit des verbes prédicatifs. L'hypothèse à vérifier est que les classes sémantiques permettraient de limiter la sur-prédiction pour ne proposer que les arguments syntaxiquement et sémantiquement possibles ;

(iii) la constitution d'une ressource réunissant toutes les catégories prédicatives. En effet, à l'heure actuelle le développement d'un programme de prédiction du contexte droit des catégories prédicatives nécessiterait de fusionner les ressources présentant les meilleures performances au regard de la tâche de prédiction envisagée, soit *LGLex*, *TreeLex* et *Nomage* (pour exemple pour les noms prédicatifs). Le projet de recherche émergent *LexVan* financé par la MESHS pour 2012-2013

vise précisément la constitution d'une telle ressource. Ce projet émergent, auquel participe la SAS Ergonotics, constitue un premier pas dans la direction que nous préconisons dans la présente étude, de nature à raccourcir, ou à tout le moins à affiner, le délai évalué pour l'élaboration d'une ressource couvrant les besoins exprimées par la SAS Ergonotics. La question de la description des catégories prédicatives au sein de ressources lexicales électroniques est donc une problématique au croisement des recherches fondamentales et appliquées, qui justifie à la fois la présente étude et de futures collaborations sur ce thème ;

(iv) de façon plus large, il est également intéressant de réfléchir sur la prédiction du contexte droit des catégories non-prédicatives comme *enfant*, qui ne possède pas de construction particulière. Une piste qui pourrait alors être explorée est celle de la notion de nom recteur définit comme étant un nom sur lequel porte une qualification de type adjectivale (ex : *enfant rougeoleux*) ou encore de type complément du nom (ex : *enfant de Max*). L'étude de Fradin (2012) va dans ce sens en tentant de déterminer la relation sémantique existante entre un adjectif suffixé par *-eux* et son nom recteur.

Références bibliographiques

Abeillé, A. (2003). *Treebanks, Building and Using Parsed Corpora*, Dordrecht : Kluwer.

Badia, T. & C. Colominas (1997). *The Predicate-Argument Structure,* Barcelona : Institut Universitari de Linguüistica Aplicada, Universitat Pompeu Fabra.

Balvet A., Barque L., Condette M.-H., Haas P., Huyghe R., Marín R. & Merlo A. (2012). La ressource Nomage : confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues*, 52/3, pp. 129-152.

Bierwisch, M. (1990-1991). "Event nominalizations : proposals and problems", *Acta Linguistica Hungarica*, 40, 1-2, pp. 19-84.

Blanche-benveniste, C. *& al.* (1984). *Pronom et syntaxe, L'approche pronominale et son application au français.* Paris. SELAF.

van den Bosch A. (2006). Scalable classification-based word prediction and confusible correction. Traitement *Automatique des Langues*, 46(2), pp :39–63.

Boissière P. & Dours D. (2001). « VITIPI : Comment un système d'assistance à l'écriture pour les personnes handicapées peut offrir des propriétés intéressantes pour le TALN ? » Actes TALN'2001, atelier TALN et Handicap, Tours. Vol. 2, p. 183-192.

Carlberger, A., J. Carlberger, T. Magnuson, M.S. Hunnicutt, S.E. Palazuelos-Cagigas & S.A. Navarro. (1997). "Profet, a new generation of word prediction: An evaluation study." Copestake, A., Langer, S. and Palazuelos-Cagigas S., editors, Natural Language Processing for Communication aids, In *Proceedings of a workshop sponsored by ACL, Madrid, Spain,* pp 23-28.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambrige, Mass.: M.I.T. Press.

Croft W. & Cruse D. A. (2004). *Cognitive Linguistics,* Cambridge : Cambridge University Press.

Danlos L. (2009). Extension de la notion de verbe support. Actes du Colloque International Supports et prédicats non verbaux dans les langues du monde, Paris.

Demonte, V. (1989). *Teoría sintáctica : de las estructuras a la rección,* Madrid : Síntesis.

Dubois, J. & F. Dubois-Charlier (1997). *Les verbes français.* Paris : Larousse.

Eynde, K. van den & P. Mertens (2006). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13/1, pp. 63-104.

Fradin, B. & Y. Yannick Mathieu (2012). Les adjectifs dérivés en -eux liés aux éléments du corps. Comment traiter des microvariations interprétatives ? Actes de 3e Congrès Mondial de Linguistique Française, volume 1, pp. 1277-1292.

Gardent, C., B. Guillaume, G. Perrier & I. Falk (2006). *Extraction d'information de sous-catégorisation à partir des tables du LADL*. Actes TALN 2006.

Garay-Vitoria N. & Abascal J. (2006). Text Prediction Systems: A survey. *Universal Access in the Information Society* (UAIS). Special Issue on "User-Centred Interaction Paradigms for Universal Access in the Information Society" (Guest editor: Christian Stary). Vol. 4, No. 3. Pp. 188-203.

Grimshaw, J. (1990). *Argument Structure*. Cambridge : MIT Press.

Grimshaw, J. & E. Williams (1993). "Nominalization and Predicative Prepositional Phrases". In Pustejovsky (éd.), pp. 97-105.

Gross, M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

Gross, G. (1989). *Les constructions converses du français*, Genève/Paris : Droz

Gross, G. (2012). *Manuel d'analyse linguistique : approche sémantico-syntaxique du lexique*, PU Septentrion : Paris.

F. Hadouche & G. Lapalme (*à par.*). *"Une version électronique du LVF comparée avec d'autres ressources lexicales"*, *Langages n° 179 -180* , Armand Colin, 2011

Hale, K. & S.J. Keyser (1986). « Some transitivity alternations in English ». Lexical Project Working Paper, 7, Centre for Cognitive Science, MIT.

van Hout, A. (1991). "Deverbal Nominalization, Object versus Event Denoting Nominals. Implications for Argument & Event Structure". In F. Drijkoningen et A. van Kemenade (éds.), *Linguistics in The Netherlands*, Amsterdam/Philadelphia : John Benjamins, pp.

Ingria, R. J. P. & G. Leland (1993). "Adjectives, Nominals, and the Status of Arguments". In Pustejovsky (éd.), pp. 107-127.

Jackendoff, R. (1983). Semantics and cognition. Cambridge, MA: MIT Press.

Jackendoff, R. (1987). "The Status of Thematic Relations in Linguistic Theory", linguistic inquiry, 18 : 3, 369-411.

Jackendoff, R. (1990). semantic structure, Cambridge : MIT Press.

Kipper-Schuler K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, June.

Kupść A**.** (2008). Adjectives in TreeLex. Dans: M. Kłopotek, A. Przepiórkowski, S. Wierzchoń et K. Trojanowski (eds.), 16th International Conference Intelligent Information Systems. Zakopane, Poland, 16-18 juin, 2008, Academic Publishing House EXIT, pp. 287--296

Kupść A. & A. Abeillé (2008). Growing TreeLex. In Gelbukh A. (éd.), *9th Int. Conf., CICLing 2008*, (Haifa, Israel, February 2008), Lecture Notes in Computational Linguistics, 4919, pp. 28-39.

Leclère, C. (2002). Organisation of the Lexicon-Grammar of French Verbs. *Linguisticæ Investigationes*, 1, vol. 25.

LEGER, C. (2006) *La complémentation phrastique des adjectifs en français.* Thèse de doctorat, Université du Québec à Montréal.

LE PÉVÉDIC (8.): 1997, *Prédiction A1orpllOsyntaxique Évoluti\'c dans ull système d'aide à la saisie de textes pOlir des personnes handicapées physiques* (111ème de doctorat en informatique, Université de Nantes).

Levin, B. (1993). *English Verb Classes and Alternation, A Preliminary Investigation.* The University of Chicago Press.

Levin, B. & M. Rappaport (1988). "Nonevent -er nominals : a probe into argument structure", linguistics, 26, pp. 1067-1083.

MacKenzie, I. S. (2002). KSPC (Keystrokes per Character) as a Characteristic of Text Entry Techniques. In *Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices*, 195-210, Heidelberg, Germany: Springer-Verlag.

Macleod, C., A. Meyers, R. Grishman, L. Barrett & R. Reeves (1998). NOMLEX: A Lexicon of Nominalizations. *Proceedings of*

EURALEX'98, Liège, août 1998.
Messiant, C., A. Korhonen & T. Poibeau (2008). Lexschem : A large subcategorization lexicon for french verbs. *Language Resources and Evaluation Conference (LREC), Marrakech*.
Meyers A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young & R. Grishman (2004). Annotating Noun Argument Structure for NomBank. *Proceedings of LREC-2004.*
Nantais, T. and F. Shein and M. Johansson (2001). "Efficacy of the word prediction algorithm in WordQTM." In *Proceedings of the 24th Annual Conference on Technology and Disability, RESNA.*
Newell A. , Stefan Langer, and Marianne Hickey. (1998). The role of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(1):1–16.
Noailly, M. (1999). *L'adjectif en français.* Paris : Ophrys.
Nunes, M. L. (1993). "Argument Linking in English Derived Nominals". In R. Van Valin (éd.), *Advances in Role and Reference Grammar,* Amsterdam/Philadelphia : John Benjamins, pp. 375-432.
Palmer M., Kingsbury P., & Gildea D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics 31(1), pp :76–105.
Pasero R. & Sabatier P. (1995). « Guided Sentences Composition : Some problems, solutions, and appli- cations », *Proceedings of NLULP'95*, Lisbonne, Portugal, p. 97-110.
Peris, A. & M. Taulé (*à par.*). Annotating the Argument Structure of Deverbal Nominalizations in Spanish.
Péry-Woodley M.-P. (1995). Quels corpus pour quels traitements automatiques ?. *Traitement Automatique des Langues*, n° 36(1-2), pp. 213-232.
Pustejovsky J. (1995). *The Generative Lexicon*, Cambridge : The MIT Press.
Rappaport, M. (1983). "On the Nature of Derived Nominals". In L. Levin, M. Rappaport et A. Zaenen (éds.), *Papers in Lexical-Functional Grammar,* Bloomington : Indiana University Linguistics Club, pp. 113-142.
Rappaport, M. & B. Levin (1988). "What to do with P-Roles". In Wilkins (éd.), pp. 7-36.
Roeper, T. (1993). Explicit syntax in the lexicon: the representation of nominalizations. *Semantics and the lexicon*, ed. by J. Pustejovsky. Dordrecht: Kluwer.
Sagot, B., L. Clément, E. Villemonte de la Clergerie & P. Boullier (2006). The Lefff 2 syntactic lexicon for French : architecture, acquisition, use. *Actes de LREC 06, Gênes, Italie.*
Sagot, B. & Danlos L. (2007). Comparaison du Lexique-Grammaire des verbes pleins et de DICOVALENCE : vers une intégration dans le Lefff. Actes de TALN'2007, Toulouse.
Saint-Dizier P., Fernandez A., Vazquez G., Kamel M. & Benamara F. (2002). The Volem Project : a Framework for the Construction of Advanced Multilingual Lexicons . In *Language Technology 2002 , Hyderabad,* , p. 123–142 : Springer Verlag, Lecture Notes. Dates de conférence : décembre 2002.
Salkoff M. & A. Valli (2005). A dictionary of french verbal complementation. Actes de Language and Technology Conference. Human Language and Technologies as a Challenge for Computer Science and Linguistics. In memory of M. Gross and A. Zampolli, Poznan, Poland.
Schadle, I. (2003). *Sibylle : Système linguistique d'aide à la communication pour les personnes handicapées.* Thèse de doctorat, Université de Bretagne-Sud.
Shein, F., T. Nantais, R. Nishiyama, C. Tam and P. Marshall. (2001). "Word cueing for persons with writing difficulties: WordQ." *The16th Annual International Conference on Technology and Persons with Disabilities, California State University at Northridge, Los Angeles, CA,* March.
Taulé, M., M. A. Marti & M. Recasens (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Proceedings of 6th International Conference on Language Resources and Evaluation, Marrakesh (Morocco).
Tenny, C. (1992). "The Aspectual Interface Hypothesis". In Sag et Szabolcsi (eds.), pp. 1-28.
Tenny, C. (1994). *Aspectual roles and the syntax-semantics interface*, Dordrecht : Foris.
Tesnière, L. (1959). *Éléments de syntaxe structurale*, Paris : Klincksieck.
Tolone, E. (2011). Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français. Thèse de doctorat, LIGM, Université Paris-Est, France.
Wandmacher T., & Antoine J.Y. (2006). Training language models without appropriate language resources: experiments with an AAC system for disabled people. Actes *LREC2006*, Genova, Italie.
Williams, E. (1981). "Argument structure and morphology", *The Linguistic Review,* 1, pp. 81-114.
Williams, E., (1987). "English as an Ergative Language : The Theta-Structure of Derived Nouns", *Proceedings of the Chicago Linguistic Society*, pp. 365-375.