Variabilité des performances des outils de TAL et genre textuel : Cas des patrons lexico-syntaxiques

Marie-Paule Jacques, Nathalie Aussenac-Gilles

Variabilité des performances des outils de TAL et genre textuel

Cas des patrons lexico-syntaxiques

Marie-Paule Jacques* — Nathalie Aussenac-Gilles**
** ERSS - UMR 5610*
Maison de la Recherche - Université de Toulouse-Le Mirail
5, Allées Antonio Machado, F-31058 Toulouse Cedex 9
marie-paule.jacques@univ-tlse2.fr

*** Institut de Recherche en Informatique de Toulouse (IRIT) - CNRS*
UPS, 118, route de Narbonne, F-31062 Toulouse Cedex
aussenac@irit.fr

RÉSUMÉ. *Nous rapportons dans cet article un ensemble de résultats liés à la mise au point d'une base de marqueurs de relations lexicales pour un outil d'aide à la réalisation d'ontologies à partir de textes,* CAMÉLÉON. *L'évaluation de ces patrons sur huit corpus différents montre une grande variation de leurs performances selon le corpus testé. Cela nous conduit à deux sortes de conclusions : 1. dans le cadre de traitements automatiques, il est nécessaire de fournir à l'utilisateur des informations sur les corpus avec lesquels ces traitements ont été mis au point ; 2. la recherche en matière de TAL doit contribuer à définir une meilleure caractérisation des textes et des genres textuels en relation avec les traitements possibles, au-delà d'une classification unique et rigide des textes.*

ABSTRACT. *In this paper, we report a set of results obtained by tuning a base of lexical relation patterns for* CAMÉLÉON, *a tool that supports ontology engineering from texts. When evaluating these patterns on eight different corpora, their efficiency varied strongly depending on the test corpus. This leads to two conclusions: 1. in the scope of NLP, users must be provided with pieces of information about the corpora used for evaluating the NLP task; 2. NLP research should contribute to a better text characterisation through an appropriate definition of text genres, which should go beyond a rigid and unique text classification.*

MOTS-CLÉS : *relations lexicales, patrons lexico-syntaxiques, genre textuel, variabilité, modélisation de connaissances.*

KEYWORDS: *lexical relations, lexico-syntactic patterns, text genre, variability, knowledge engineering.*

1. Introduction

La question que nous abordons dans cet article est celle de la variabilité des résultats d'un même traitement appliqué à différents genres de textes. La plupart du temps, les recherches en matière de traitement automatique, lorsqu'elles présentent des analyses sur corpus, indiquent des résultats obtenus sur seulement UN corpus ou un ensemble indistinct de textes. Il est vrai que pour de nombreuses tâches, les premiers problèmes à résoudre ont d'abord été de déterminer comment réaliser ces tâches et comment les réaliser de façon optimale avant de mesurer si ce comment devait ou pouvait être adapté au type de texte à traiter. Mais maintenant qu'une certaine maturité a été atteinte et que l'on cerne mieux comment réaliser tel ou tel type d'analyse, il est possible de faire entrer en jeu les nuances requises par la diversité des textes. On voit d'ailleurs émerger cette problématique dans des travaux récents : C. Frérot (2005) montre dans sa thèse que les performances de différentes stratégies de rattachement prépositionnel (pour une tâche d'analyse syntaxique) sont variables selon le genre de corpus considéré ; dans le domaine de la recherche d'informations, la difficulté des requêtes peut être calculée et prédite à partir de l'analyse d'un certain nombre de traits linguistiques (Mothe et Tanguy, 2005) ; nous évoquerons en fin d'article différents travaux qui, s'inspirant des propositions de D. Biber (1988) pour l'anglais, mettent au point des outils de typologie textuelle.

Nous apportons ici une pièce supplémentaire en faveur d'une conception diversifiée et adaptative des traitements automatiques. L'étude dont nous présentons les résultats s'inscrit dans le cadre de l'amélioration d'un outil d'aide à l'élaboration de terminologies et d'ontologies à partir de textes, CAMÉLÉON. Nous nous focalisons ici sur un module spécifique de CAMÉLÉON, la base générique de patrons lexico-syntaxiques fournie à l'utilisateur avec l'outil et destinée à lui permettre d'amorcer la constitution d'une terminologie et/ou d'une ontologie sur son propre corpus. Dans l'esprit initial de CAMÉLÉON, cette base était destinée à stocker des patrons éprouvés, c'est-à-dire donnant de bons résultats en termes de performance sur n'importe quel texte. Afin de déterminer quels patrons choisir pour constituer cette base, nous avons procédé à une évaluation sur différents corpus. Cette évaluation a révélé de très grandes différences de performances des patrons selon le corpus de textes sur lequel ils ont été testés, ce qui remet en cause leur généricité. Cela nous conduit à deux sortes de conclusions : 1. pour la mise à disposition d'un outil qui comporte un ou plusieurs modules de traitements automatiques, il est nécessaire de fournir à l'utilisateur des informations sur la façon dont ces traitements ont été mis au point, c'est-à-dire des informations qui permettront à l'utilisateur de cerner précisément les limites de la validité de ces traitements (par exemple, une même tâche peut être très bien réussie sur les textes du journal Le Monde et très médiocrement sur des constats d'accidents) ; 2. de manière plus globale, la recherche en matière de TAL doit intégrer la question des genres textuels, même de façon limitée et très basique.

Dans la partie qui suit, nous indiquons l'architecture générale du logiciel CAMÉLÉON et la place de la base générique de patrons, ce afin de situer les enjeux

plus spécifiques de la constitution de cette base. Dans la partie 3, nous décrivons le processus de mise au point des patrons et nous en indiquons les résultats section 4. Les parties 5 et 6 sont consacrées à une discussion de ces résultats et à la conclusion.

2. Aperçu de CAMÉLÉON

CAMÉLÉON est un outil d'aide au repérage de relations conceptuelles à partir d'analyse de textes (Séguéla, 2001). C'est donc un logiciel de traitement du langage défini pour être intégré dans une démarche supervisée d'ingénierie des connaissances. Plus précisément, CAMÉLÉON permet d'élaborer, à partir de l'analyse de corpus écrits, des réseaux conceptuels à composante terminologique, pouvant servir de base à la définition de terminologies ou d'ontologies. Pour cela, CAMÉLÉON propose une approche par marqueurs, faisant l'hypothèse que les relations lexicales peuvent fournir des indices pour définir des relations conceptuelles, et avec elles, de nouveaux concepts et termes associés.

Nous présentons en 2.1 la démarche générale qui englobe l'utilisation de CAMÉLÉON, et précisons l'aide fournie par le logiciel à chacune des étapes de la démarche. Dans la partie 2.2, nous nous focalisons sur le processus de définition et d'utilisation des patrons, qui correspond à une première étape. En effet, le travail de mise au point et d'évaluation de patrons rapporté dans cet article est exemplaire de cette première étape. De plus, le fruit de ce travail, la « base générique de patrons » devrait faciliter le processus de recherche de relations pour un nouveau projet. Nous exposons rapidement en 2.3 la manière dont les patrons sont ensuite utilisés pour enrichir un modèle conceptuel.

2.1. *Démarche générale pour l'identification de relations*

La démarche préconisée par CAMÉLÉON consiste à enrichir un réseau conceptuel en se focalisant sur le repérage de relations, et cela à partir de régularités de forme relevées dans des textes à l'aide de marqueurs. Ces marqueurs correspondent à des patrons[1] comportant des éléments lexicaux et syntaxiques. En effet, la dernière version du logiciel[2] suppose que le corpus de textes ait été étiqueté grammaticalement[3]. L'approche par patrons fait l'hypothèse que l'interprétation de ces éléments de forme définit régulièrement le même rapport de sens entre les

[1] Par la suite de la partie 2, nous utiliserons indifféremment les termes « marqueur » et « patron », bien qu'ils puissent recouvrir chacun une réalité différente dans d'autres contextes.

[2] CAMÉLÉON est un logiciel développé en Java. Il utilise MySQL pour gérer et enregistrer les corpus, marqueurs et modèles définis, et un concordancier, KESKYA, développé à l'aide de modules de la suite logicielle Emdros, pour interroger les bases de textes. Il permet d'exporter des ontologies à composante terminologique.

[3] Actuellement, l'étiqueteur utilisé est le système TreeTagger, pour le français ou l'anglais. Mais tout autre étiqueteur peut être utilisé si l'on fournit le jeu d'étiquettes en paramètre. http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/

termes. La projection des patrons sur le corpus met en avant des fragments de texte dont l'interprétation peut ensuite donner lieu, dans le réseau conceptuel, à la définition de nouveaux termes, concepts ou relations entre concepts, les relations contribuant à définir les concepts (Aussenac-Gilles et Séguéla, 2000). Pour un projet donné, c'est-à-dire un corpus de textes et un objectif de modélisation particuliers, la démarche préconisée par CAMÉLÉON se déroule donc en deux grandes étapes :

1) mettre au point un jeu de marqueurs adaptés à ce projet, ce qui suppose de projeter des patrons qui seront ou non retenus, et d'en évaluer quelques occurrences ;

2) projeter ces marqueurs sur le corpus, en interpréter toutes les occurrences et étendre ou corriger le modèle conceptuel en conséquence.

Le logiciel offre donc deux ensembles de fonctionnalités adaptées à chacune de ces étapes. Le caractère non automatisable de l'interprétation des phrases à chaque étape donne à l'utilisateur (cogniticien ou linguiste) un rôle primordial.

Pour faciliter son travail, et en conformité avec des travaux linguistiques sur les relations sémantiques (Jackiewicz, 1996) (Garcia, 1998) (Condamines et Rebeyrolle, 2000), un ensemble de patrons, la « base générique de marqueurs », est proposé à l'utilisateur. Cette base est destinée à fournir des marqueurs « prêts à l'emploi » pour amorcer le processus. Bien que la base soit qualifiée de générique, ces marqueurs peuvent ne pas fonctionner sur tout type de corpus (Condamines, 2002), ou encore ils peuvent être plus efficaces après leur adaptation aux caractéristiques du corpus de textes analysés (Rebeyrolle et Tanguy, 2000) (Hamon et Nazarenko, 2001).

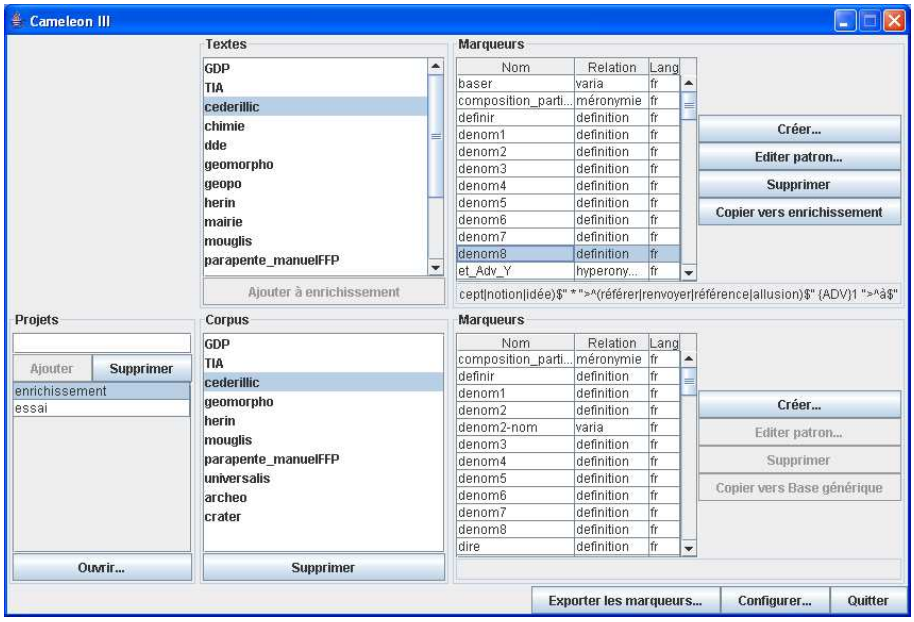

Figure 1: *Fenêtre principale de CAMÉLÉON.*

CAMÉLÉON prévoit donc que les marqueurs puissent être adaptés pour chaque projet et chaque corpus. La fenêtre d'accueil de CAMÉLÉON (figure 1) permet de définir un projet, d'y associer un corpus et de constituer une liste de marqueurs pour ce projet (liste inférieure sur la figure 1). Pour cela, les marqueurs peuvent être sélectionnés à partir de la base générique de marqueurs (liste supérieure) et adaptés, ou bien définis à l'aide de l'éditeur de marqueurs (auquel on accède via le bouton « créer »).

2.2. *Mise au point de patrons*

La mise au point d'un jeu de patrons pour un projet donné est un processus cyclique qui, pour chaque patron, alterne sa modification et son évaluation par projection sur le corpus.

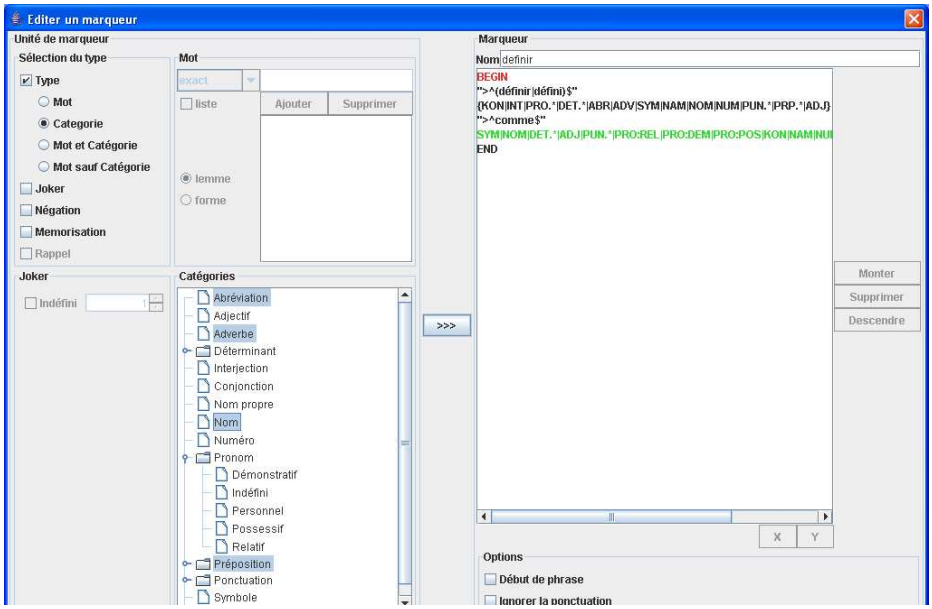

Figure 2. *L'éditeur de patrons de CAMÉLÉON.*

Pour définir de nouveaux marqueurs ou modifier des marqueurs sélectionnés dans la base générique, l'utilisateur utilise l'éditeur du concordancier KESKYA qui fait partie de CAMÉLÉON (figure 2). Il doit caractériser les éléments lexico-syntaxiques présents avant, après et entre les termes en relation. Pour cela, il peut fournir un terme (zone Mot sur la partie gauche de la figure 2) et/ou une étiquette morpho-syntaxique parmi celles utilisées par l'étiqueteur (à choisir dans la liste de catégories sur la partie gauche, figure 2). L'utilisateur peut aussi caractériser la forme des éléments mis en relation par le marqueur, appelés X et Y. Ces éléments

sont mis en évidence par une coloration dans le patron (affiché à droite de la figure). Par exemple, sur la figure 2, le marqueur édité (*définir*) permet de rechercher des formes comme “X est défini comme Y”. L’utilisateur n’a pas précisé la nature de X (BEGIN est donc coloré comme étant X). La liste colorée avant END contraint la formulation de Y.

L’évaluation d’un marqueur passe par la lecture d’une partie de ses occurrences, c’est-à-dire de phrases qu’il permet de repérer dans chacun des textes du corpus (partie gauche de la figure 3). Chaque phrase peut être visualisée séparément (partie inférieure de la figure 3). Le logiciel colorie dans la phrase les parties reconnues par le marqueur comme étant les éléments mis en relation (X et Y). On vérifie que le marqueur présente dans les phrases un sens stable, qui est bien celui de la relation associée. La comptabilisation des phrases examinées et validées contribue à calculer la précision du marqueur (en haut à droite de la figure 3). Ces différents éléments permettent à l’utilisateur de décider s’il conserve ou rejette le marqueur (boutons *valide/invalide* sur la figure 3), s’il en modifie la forme ou le type de relation associé (bouton *editer patron*).

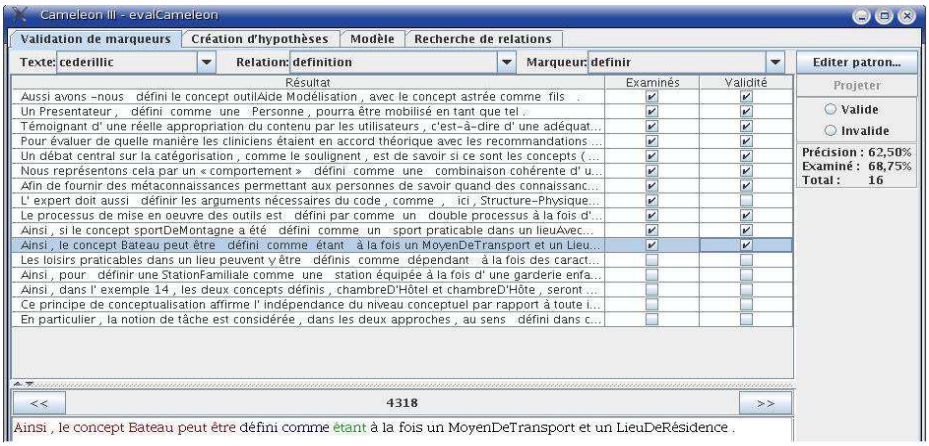

Figure 3: *Evaluation d’un patron avec CAMÉLÉON*.

2.3 *Enrichissement d’un modèle conceptuel*

Une fois stabilisé un ensemble de patrons et de relations, l’enrichissement du modèle se déroule en deux temps. Tout d’abord, via l’interface présentée en figure 4, l’utilisateur dépouille systématiquement toutes les occurrences des marqueurs projetés pour retenir des hypothèses de relation. À ce niveau, si une occurrence est pertinente, il sélectionne des termes mis en relation (partie en bas à gauche de la figure 4) et le type de cette relation. Ensuite, l’utilisateur intègre ces éléments dans le modèle en cours de construction. Pour cela, il doit associer les termes à des concepts, et décider des concepts qui vont être mis en relation.

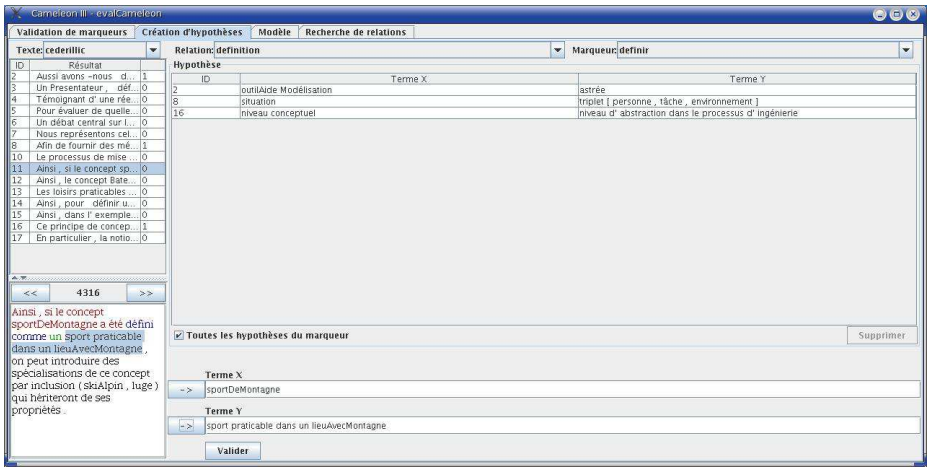

Figure 4 *: Extraction d'hypothèses de relations.*

Au cours de ces deux phases, l'utilisateur interprète chaque phrase en tenant compte de son contexte en corpus, mais aussi des objectifs de modélisation. Étant donné l'occurrence d'un marqueur, l'utilisateur se demande si les termes utilisés, les concepts qu'ils peuvent évoquer et les relations identifiées entre ces concepts sont ou non pertinents pour le modèle en cours de construction. Le choix des concepts et relations qui vont représenter au mieux, en cohérence avec le modèle, ces connaissances n'est également pas trivial. Les concepts à mettre en relation ne sont pas nécessairement ceux exprimés dans la phrase considérée. On peut préférer définir la relation entre des concepts parents ou fils de ceux mentionnés.

3. Evaluation de la base de patrons

Le processus proposé par CAMÉLÉON repose donc sur une première étape consistant à extraire de corpus de textes des zones supposées comprendre l'expression d'une relation lexicale. Cet article n'a pas pour objet de discuter du bien-fondé des principes de fonctionnement de CAMÉLÉON, argumentés dans (Aussenac et Séguéla, 2000) et (Séguéla, 2001). Depuis que les travaux en terminologie et dans le domaine des ontologies mettent en avant l'idée que l'on peut acquérir des connaissances à partir de textes, l'extraction de zones de textes pour identifier des relations conceptuelles est une pratique plutôt triviale. Elle repose sur l'hypothèse qu'il existe des moyens stables d'exprimer telle ou telle relation. Cependant, cette idée est si bien établie que l'analyse et la mesure d'une variation, pourtant intuitivement perçue par les analystes, ont été reléguées à l'arrière-plan. La myopie des analyses automatiques s'explique par deux facteurs qui méritent que l'on s'y attarde quelque peu : l'historique du développement des travaux dans ce domaine et leur dimension très applicative.

3.1. *Marqueurs, patrons : une idée de stabilité*

Quoique l'on pourrait certainement faire remonter plus loin l'origine de la notion de marqueur de relation[4], on peut considérer les propositions de M. Hearst (1992) comme une étude fondatrice qui inspira de nombreux travaux en TAL et en terminologie. L'objectif principal de la démarche est de projeter sur un texte ou un corpus de texte, un marqueur que l'on définira comme une « *forme linguistique faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques) dont l'interprétation définit régulièrement le même rapport de sens entre les termes* » (Haddad, 2002 : 37) ou, d'une façon plus élaborée, un patron lexico-syntaxique : *« à la différence des marqueurs, les patrons identifient la relation recherchée plus précisément en définissant également des contraintes syntaxiques ou typographiques sur le contexte des termes »* (Grabar et Hamon, 2004 : 72). Quelle que soit la forme de ces expressions projetées sur le corpus, l'idée sous-jacente est qu'elles expriment une relation donnée avec suffisamment de stabilité pour produire des résultats acceptables en termes de rappel et de précision. C'est à peu près ce que M. Hearst écrit à propos des marqueurs de la relation d'hyperonymie qui sont à la base de son travail :

« (i) They occur frequently and in many text genres.
(ii) They (almost) always indicate the relation of interest.
(iii) They can be recognized with little or no pre-encoded knowledge.

Item (i) indicates that the pattern will result in the discovery of many instances of the relation, item (ii) that the information extracted will not be erroneous, and item (iii) that making use of the pattern does not require the tools that it is intended to help build. » (Hearst, 1992: 540).

On notera surtout les deux premiers items qui manifestent un certain credo de généralité et de généricité. Ce credo a, dès le départ, imprégné les recherches en matière de marqueurs de relations (par le terme de *marqueur*, nous englobons aussi les patrons lexico-syntaxiques). Mais bien sûr, les chercheurs ne sont pas naïfs et ignorants des faits de langue, tout le monde est parfaitement conscient que selon le corpus analysé, les résultats pourront être variables. Cela n'a cependant pas donné lieu à une étude précise de cette variation. La littérature envisage généralement les relations lexicales et leurs marqueurs comme se répartissant en deux sous-ensembles : des relations indépendantes du domaine, l'hyperonymie et la relation partie-tout, dont l'expression a globalement été bien décrite, voir par exemple pour le français (Borillo, 1996) et (Jackiewicz, 1996) ; des relations « spécifiques » ou « transversales », propres à un domaine, qu'il faut décrire au coup par coup. S'il paraît évident que les marqueurs des secondes sont difficilement exportables et réutilisables dans un corpus différent de celui dans lequel ils ont été décrits, les premières sont supposées offrir une certaine stabilité et si leur repérage dans les

[4] E. Morin (1999) évoque Robison, Harold (1970): Computer-detectable semantic structures, *Information Storage and Retrieval*, vol. 6, pp. 273-288.

textes présente des différences en termes de résultats, ces différences sont généralement négligées. Cela tient, nous semble-t-il, à un second facteur, lié à la tâche elle-même.

Hormis l'étude de Hearst et les travaux de linguistes cités plus haut, la plupart des recherches en matière de repérage de relations lexicales sont liées à une application précise : il ne s'agit pas d'inventorier tous les modes d'expression des relations X ou Y pour la beauté de la science, il s'agit d'élaborer effectivement une terminologie structurée ou une ontologie d'**un** domaine. La perspective est clairement opérationnelle. La question de la variation des résultats devient alors secondaire au profit de la définition d'une méthode reproductible : comment procéder ? comment éliminer le bruit ? quelles ressources utiliser (thésaurus, dictionnaires, terminologie préexistante, etc.) ? Ce qui est destiné à enrichir le savoir collectif est moins le patron lui-même que la façon de l'acquérir. Même s'il peut paraître intéressant d'augmenter une liste de marqueurs, le marqueur ou le patron lexico-syntaxique restent anecdotiques, d'autant que l'objectif est d'ouvrir la voie vers les relations dites transversales que nous évoquions au paragraphe précédent. De nombreux travaux se penchent explicitement sur ces relations, par exemple (Girju et Moldovan, 2002 ; Maedche et Staab, 2000 ; Riloff, 1996 ; Tchalakova *et al.*, à paraître ; Yamaguchi, 2001). Cependant, on ne trouve dans ces articles ni inventaire de marqueurs, ni liste de patrons ; leur perspective est autre : ils visent la définition du comment faire[5].

3.2. *Mise au point de la base de patrons*

La base de marqueurs « génériques » de CAMÉLÉON est destinée à stocker et à fournir à l'utilisateur un certain nombre de patrons « prêts à l'emploi ». A lui ensuite d'adapter ces patrons, de les modifier et de les multiplier selon le corpus et le domaine sur lequel il travaille. Notre objectif initial était de remplir la base à l'aide de patrons éprouvés, suffisamment généraux (au sens indiqué plus haut : que l'on trouve dans beaucoup de textes et qui signalent – presque – toujours la relation visée) pour être d'une véritable utilité sur tout corpus de textes.

Les patrons que nous avons entrés dans cette base proviennent de différentes sources : la version précédente de CAMÉLÉON, elle-même fondée sur toute l'analyse de son auteur (Séguéla, 2001), des travaux de recherche antérieurs[6] et, de manière

[5] La « bipolarité » que nous venons d'esquisser correspond à peu près aux deux points de vue évoqués par Adeline Nazarenko en ouverture d'une table ronde consacrée aux marqueurs lors de la Semaine de la Connaissance 2006 : d'un côté, une description « guidée par les données » ; de l'autre, une approche « guidée par le but ».

[6] Nous remercions Josette Rebeyrolle et Ludovic Tanguy pour nous avoir prêté les corpus sur lesquels ils ont testé leur repérage automatique de définitions (Rebeyrolle et Tanguy, 2001) et pour nous avoir aussi donné les différents patrons produits lors cette étude.

plus générale, la littérature sur l'expression des relations lexicales, plus précisément les études qui se consacrent à la description de formes linguistiques. Au total 70 patrons ont été définis et évalués : 18 pour le repérage des définitions, 35 pour la relation d'hyperonymie, 14 pour la méronymie, 1 pour les reformulations et 2 divers.

Chaque patron a été saisi à l'aide de l'outil d'élaboration de patrons de CAMÉLÉON. Les patrons de définition fournis par Rebeyrolle et Tanguy n'ont pas subi de modification ; les autres ont été mis au point de façon relativement *ad hoc*. Dans un premier temps, le patron a été entré « tel quel », puis projeté sur les différents corpus et, dans un second temps, modifié ou affiné en fonction des contextes obtenus, de façon à éliminer un maximum de bruit sans réduire le nombre de contextes pertinents.

Un exemple de modification tient au fait que les patrons issus de la précédente version de CAMÉLÉON ne font pas usage d'étiquettes de catégorie grammaticale. Dans certains cas, nous avons remplacé des « jokers » (un mot, deux mots, etc.) par des contraintes sur la catégorie grammaticale de l'élément autorisé. Les modifications ont été motivées par ce que nous observions des contextes renvoyés par l'outil et non par un quelconque sentiment linguistique sur la construction en question, c'est en ce sens que nous les jugeons *ad hoc*[7].

Pour donner un exemple concret, le patron « le Y être le X le plus|moins … » (où Y représente l'hyponyme et X l'hyperonyme) permet de récupérer le contexte visé (nous mettons en gras la partie de l'énoncé correspondant au patron) :

*La lave des coulées **est** la roche volcanique **la plus** résistante.*

Mais dans certains corpus, cette formulation est employée de façon régulière comme forme d'insistance, par exemple :

*La méthode KOD en **est** l'exemple **le plus** frappant*

On a ainsi, en position « le X le plus », *l'exemple le plus|moins*, *le cas le plus|moins*, *le résultat le plus|moins* et il ne semble pas impossible d'accroître la liste. Pour éviter ces contextes qui ne sont pas susceptibles de fournir des éléments de l'ontologie, on a donc stipulé que les noms *exemples*, *cas*, *résultat* ne doivent pas figurer à la place réservée à l'hyperonyme. D'autres corpus pourraient conduire à ajouter d'autres termes ou au contraire à lever totalement ou partiellement cette contrainte, en fonction du projet qui guide l'analyse.

3.3 *Corpus et méthode*

L'évaluation que nous fournissons est une mesure faite au terme de ce processus de mise au point. Dans la perspective qui est la nôtre, celle de l'élaboration d'un

[7] (Tchalakova *et al.*, 2006) font aussi état de cette nécessité de tester et affiner les patrons sur corpus afin d'en définir une formulation satisfaisante.

outil d'aide à la construction d'ontologies, nous devons non seulement fournir des listes de marqueurs et de patrons, mais encore avoir une idée aussi correcte que possible de leur intérêt pour le futur utilisateur. Nous avons donc décidé de tester les patrons de la base générique de CAMÉLÉON sur différents corpus, huit au total :

1. un guide de planification de réseau électrique (GDP, 187 800 mots) ;

2. des articles scientifiques de la conférence Ingénierie des Connaissances (IC, 198 500 mots) ;

3. un manuel de géomorphologie (GEO, 260 000 mots) ;

4. un manuel de spécification de logiciels dans le domaine de l'électricité (MOU, 57 500 mots);

5. des articles extraits de l'Encyclopedia Universalis, du domaine de la géomorphologie (ENC, 200 500 mots) ;

6. un manuel de parapente (PAR, 23 800 mots) ;

7. plusieurs thèses en archéologie (ARCH, 95 000 mots) ;

8. des textes du domaine de la télécommunication[8] (CRAT, 1 000 000 mots).

Précisons que le choix de ces corpus ne correspond à aucune hypothèse préalable, il est plutôt « opportuniste » dans le sens où nous souhaitions mettre au point et évaluer les patrons sur un maximum de textes différents et que nous avons donc utilisé les corpus que nous avions à notre disposition suite à différents projets de recherche.

4. Résultats

Nous présentons ici les résultats obtenus pour seulement un échantillon (pour des raisons de place).

4.1. *Modalités d'évaluation*

Pour une partie seulement des patrons testés, les patrons de définition, nous avions une liste de référence, c'est-à-dire la liste de toutes les phrases des corpus que chaque patron devait retrouver. Ceci nous a permis, pour chaque patron, d'avoir une mesure de rappel et une mesure de précision. Nous avons ainsi pu vérifier le rappel sur ce sous-ensemble. Pour tous les autres patrons, nous n'avions pas de liste de référence, donc pas de moyen de vérifier le rappel. Nous n'avons alors pour l'évaluation que la mesure de précision calculée automatiquement par le logiciel, seule mesure mentionnée dans les tableaux de résultats fournis par la suite.

[8] Ceux qui forment le corpus CRATER www.comp.lancs.ac.uk/ucrel/corpora.html#crater

En conséquence, nous n'avons presque pas étudié les variations en termes de rappel. Nous ne pouvons qu'extrapoler à l'ensemble des patrons et des corpus ce que nous avons constaté sur les 18 patrons de définition : le rappel varie aussi, parfois considérablement selon le patron, comme le montre le tableau 1. Pour chaque colonne de ce tableau, les deux premières lignes indiquent une fourchette de variation du taux de rappel, la dernière ligne précise le nombre de patrons concernés par cette fourchette.

Min.	100	50	80	88	91	96	98	82	0	0
Max.	100	100	100	100	100	100	100	86	54	93
Nbre	8	1	1	1	1	1	1	1	1	1

Tableau 1. *Fourchettes de variation du rappel pour les patrons de définition.*

On peut déjà tirer un enseignement, connu des statisticiens et qui ne va faire que se confirmer au fur et à mesure de l'analyse des résultats : une mesure moyennée, comme celles que l'on trouve souvent, masque une grande disparité de performances. Voyons maintenant l'ensemble des résultats.

4.2. *Evaluation*

Au niveau de l'évaluation, nous avons distingué deux sous-ensembles de patrons : les patrons de définition, que nous n'avons testés que sur les cinq corpus pour lesquels nous avions une liste de référence ; les autres patrons (hyperonymie, méronymie, reformulation et divers), que nous avons testés sur tous les corpus. C'est pourquoi nous fournirons les résultats dans deux tableaux différents. Nous rappelons qu'il ne s'agit là que d'un échantillon, on trouvera en annexe les patrons et, pour chacun, les valeurs extrêmes de leur taux de précision sur les différents corpus.

Les patrons de définition. Ils ont pour objectif de retrouver dans les textes les énoncés définitoires, c'est-à-dire les énoncés par lesquels l'auteur d'un texte explicite le sens des mots qu'il emploie.

Dans le tableau 2, nous ne donnons le nombre de contextes renvoyés par chaque patron qu'à titre indicatif, pour permettre d'apprécier à sa juste valeur la mesure de précision qui autorise une réelle évaluation du patron. Il nous semble en effet que deux taux de précision identiques n'ont en fait pas la même valeur selon le nombre d'occurrences sur lequel ils sont calculés : une précision de 14% obtenue sur 7 occurrences n'est en fait guère comparable à une précision de 15% obtenue sur 375 occurrences, ne serait-ce que parce que la mise au point du patron peut plus facilement s'appuyer sur des régularités linguistiques avec 375 occurrences qu'avec 7. Mais ce nombre de contextes ne peut guère permettre une comparaison entre les corpus dans la mesure où ceux-ci ne sont pas de même taille.

	GDP		IC		GEO		MOU		ENC	
	N	*P*	N	*P*	N	*P*	N	*P*	N	*P*
définir	3	*100*	43	*98*	0		2	*100*	2	*100*
dénom2	7	*29*	10	*10*	57	*89*	0		23	*100*
entendre par	0		7	*71*	3	*33*	2	*100*	0	
signifier	0		13	*38*	29	*76*	0		7	*14*

Tableau 2. *Extrait des résultats des patrons de définition (N = nombre de contextes retrouvés par le patron, P = taux de précision du patron exprimé en pourcentage)*

Le tableau 2 montre de façon assez évidente qu'il n'y a guère que le patron *définir* à présenter un taux de précision plutôt stable et apparemment peu dépendant du corpus. Ce patron est le suivant (les étiquettes morphosyntaxiques sont celles du TreeTagger, voir note 3) :

">^(définir|défini)$"
{NAM|PRO.*|DET.*|NUM|PRP.*|ABR|ADV|NOM|SYM|KON|ADJ|PUN.*|INT}*
">^comme$"
KON|NOM|ADJ|PRO:POS|PRO:DEM|SYM|ABR|PRO:REL|NUM|VER.*|INT|NAM|PUN.*|DET.*|PRO:IND

Le signe > avant une forme indique qu'il s'agit d'un lemme, les signes ^ et $ de part et d'autre de la forme indiquent qu'il s'agit d'un mot complet (et non d'une partie de mot), la barre verticale | indique que les éléments sont optionnels, les accolades { } délimitent un ensemble d'éléments, l'astérisque * 1. après cet ensemble indique que les éléments peuvent se répéter un nombre indéfini de fois, 2. à l'intérieur d'une étiquette de catégorie permet de recouvrir n'importe quel caractère. Le patron ci-dessus peut donc être formulé : le lemme *définir* ou le lemme *défini*, suivi d'un nombre indéfini d'occurrences de l'une des catégories grammaticales Nom Propre ou Pronom ou Déterminant ou Nombre ou Préposition ou Abréviation ou Adverbe ou Nom Commun ou Symbole ou Conjonction ou Ponctuation ou Interjection, suivi du lemme *comme* suivi d'une unité de catégorie grammaticale Conjonction ou Nom Commun ou Adjectif ou Pronom Possessif ou Pronom Démonstratif ou Symbole ou Abréviation ou Pronom Relatif ou Nombre ou Verbe ou Interjection ou Nom Propre ou Ponctuation ou Déterminant ou Pronom Indéfini.

Ce patron permet de retrouver des contextes tels que :

*Un Projet Logiciel peut se **définir comme** un Processus de Développement.*

*La période d'actualisation /0, T/ est **définie comme** la période de temps sur laquelle s'échelonnent les dépenses dont on veut connaître le total actualisé.*

La stabilité de ce patron s'explique par le caractère très explicite de la définition dans ce cas : si l'auteur emploie *définir*, c'est en général pour … définir. Par contre, il ressort que la fiabilité des autres patrons en tant qu'expression d'une définition est très dépendante du corpus. Par exemple, le patron *dénom2* décrit une construction

moins univoquement consacrée à une définition. Il combine les verbes *porter appliquer employer réserver recevoir prendre utiliser donner proposer mériter* avec des mots-clés tels que *nom terme mot expression vocable appellation désignation dénomination*, ce qui permet de retrouver des contextes tels que :

Dans les systèmes rhexistasiques caractérisés par la semi-aridité ou le rôle du gel, les versants dus à un affleurement rocheux résistant ont eux -mêmes un profil convexe en haut et concave en bas ; mais le rocher affleure, non recouvert par le manteau de débris ; la convexité sommitale a ***reçu le nom de*** *waxing slope (pente croissante), l'abrupt rocheux est la free face ;*

Pour comparer ces stratégies on ***utilisera les termes classiques de*** *la valorisation (pertes, défaillance, gains d'exploitation).*

On voit clairement que le premier est bien une définition qui indique que *waxing slope* est *la convexité sommitale* [du versant], alors que le second ne définit pas un mot mais permet à l'auteur de préconiser l'emploi d'une certaine terminologie dans certains cas (*pour comparer ces stratégies*). Il ne semble pas utile de multiplier les exemples pour la conclusion à laquelle conduit la suite de l'évaluation.

Les autres patrons. Le tableau 3, qui indique aussi nombre d'occurrences et précision pour quelques autres patrons, manifeste exactement les mêmes tendances : la variation qui touche les performances des patrons parfois se limite à quelques points (par exemple *inclure*), parfois s'échelonne de 0 à 100. Or, nous ne donnons pas ici les résultats de patrons « exotiques » mais bien de ce que la littérature linguistique et « TAL » a noté comme moyens réguliers d'exprimer la relation concernée : hyperonymie pour les trois premiers, méronymie pour les deux suivants, localisation et reformulation pour les deux derniers.

	GDP		**IC**		**GEO**		**MOU**	
	N	*P*	N	*P*	N	*P*	N	*P*
être-un	268	*19*	574	*19*	752	*23*	129	*12*
et Adv	10	*10*	15	*7*	56	*30*	6	*17*
sorte de	0		7	*57*	3	*67*	0	
inclure	75	*51*	32	*41*	16	*50*	18	*61*
partie de	0		0		7	*0*	0	
situé dans	40	*53*	63	*38*	38	*24*	4	*50*
c'est-à-dire	6	*67*	37	*54*	40	*80*	3	*100*
	ENC		**PAR**		**ARCH**		**CRAT**	
	N	*P*	N	*P*	N	*P*	N	*P*
être-un	420	*16*	62	*40*	181	*29*		
et Adv	66	*5*	2	*0*	13	*38*	19	*58*
sorte de	1	*100*	0		0		4	*100*
inclure	29	*62*	2	*100*	27	*19*	267	*48*
partie de	1	*100*	1	*0*	1	*0*	11	*18*
situé dans	55	*24*	4	*75*	36	*56*	291	*59*
c'est-à-dire	14	*29*	2	*100*	8	*63*	11	*64*

Tableau 3. *Résultats pour quelques autres patrons*

Voici des exemples de contextes visés par chacun de ces patrons.

- est-un

*L'albédo d'un corps **est un** rapport qui exprime la partie de rayonnement directement réfléchie et donc non absorbée.*

- et Adv (Adv = *notamment, notablement, spécialement, particulièrement*)

*En ce qui concerne les grandes stations **et particulièrement** les stations Intelsat de type A…*

- sorte de

*les amines, qui sont des **sortes de** substances chimiques ;*

- inclure

*Les services de base à fournir dans le Rmtp **comprennent** les téléservices et les services support…*

- partie de

*l'ontologie **est un composant de** la mémoire d'entreprise…*

- situé dans

*Le Ccm interroge l'Elv à chaque fois qu'il a besoin d'informations relatives à une station mobile donnée **située** à ce moment **dans** la zone du Ccm.*

- c'est-à-dire

*la résolution, **c'est-à-dire** la taille des objets qui se distinguent, est de 100 m.*

Ce que nous voulons mettre en discussion ici, ce n'est pas la faiblesse de la précision de certains de ces patrons – il est bien connu que, par exemple, *être-un* génère énormément de bruit – c'est l'importante variation de cette précision selon les corpus. Pour rester avec l'exemple de *être-un* qui fournit une grande quantité de contextes et, par là, une mesure valide, on observe trois groupes de valeurs : autour de 10-15%, autour de 20-30% et un score inhabituel de 40% dans un des corpus. De tels résultats suscitent diverses remarques.

5. Discussion

La différence de performances d'un même patron, illustrée par l'échantillon de résultats fournis dans les tableaux 2 et 3, a, selon nous, quatre conséquences majeures qui sont généralement peu prises en compte dans les travaux en TAL.

1. Par rapport à la dépendance au corpus. Lorsque l'on élabore des traitements automatiques en prenant en compte des données linguistiques réelles, en d'autres termes en travaillant à partir de corpus, une fiabilité identique du traitement élaboré n'est absolument pas garantie dès lors qu'on l'applique sur un corpus de genre différent de celui qui a servi de corpus d'analyse. C'est là un fait bien connu de nombre des chercheurs travaillant sur corpus. (Habert *et al.*, 2000) rapportent diverses expériences dans diverses tâches (étiquetage morpho-syntaxique, analyse

syntaxique, recherche d'information) montrant une faible transportabilité des traitements. Pour la tâche qui nous préoccupe, l'identification de marqueurs et patrons lexico-syntaxiques pour repérer les relations conceptuelles dans les textes, cela implique une dépendance cruciale au corpus d'analyse. En effet, la validité d'un patron est généralement déterminée par sa performance, i.e. sa capacité à retrouver toutes les occurrences de la relation visée (ou le plus grand nombre) sans noyer l'analyste dans un flot de contextes non pertinents. Si l'analyste privilégie la précision par rapport au rappel, tel patron pourra être écarté précisément parce qu'il « ramène » trop de bruit – ce que fait par exemple Morin (1999) avec le patron "est-un". Le problème que montre notre étude est qu'en fait, la performance peut être divisée ou multipliée par deux ou plus selon les textes utilisés et que la décision de conserver ou d'écarter un patron pourra être purement et simplement inversée selon le corpus sur lequel on teste ce patron. Par exemple, le patron que nous avons appelé *« et Adv »*, issu de (Borillo, 1996), serait vraisemblablement rejeté comme trop bruité par un analyste qui le testerait uniquement sur le corpus IC (7% de précision), mais pourrait être considéré plus favorablement par un analyste qui le testerait sur le corpus CRATER (58% de précision). Cela pose alors la question de ce que l'on entend par *général* ou *spécifique*.

2. Par rapport à la notion de généricité. Nous avons souligné dans la section 3.1 l'opposition entre deux types de relations : des relations telles que l'hyperonymie et la relation partie-tout qui sont les éléments structurants des ontologies et que l'on s'attend à retrouver dans tous les domaines ; des relations spécifiques, dépendantes d'un domaine de la connaissance, que l'on pourra trouver ici mais pas là. Entrent dans cette dernière catégorie des relations comme *a pour symptôme* limitée au domaine de la médecine ou *est sous-traitant de* que nous avions identifiée lors d'une étude dans le domaine des transports en communs (Jacques et Soubeille, 2000).

La présente étude tend à montrer que si les premières peuvent à juste titre être considérées comme relativement générales et assez peu dépendantes du domaine, il faut se garder d'étendre ce caractère de généralité à leur expression. En d'autres termes, si généralité il y a, elle concerne la relation – effectivement les entités du domaine sont conceptualisées comme étant plus spécifiques ou plus génériques, certaines comme des touts, d'autres des parties, etc. – **et non le marqueur de cette relation**. Parmi tous les moyens que la langue met à disposition pour l'expression de telle ou telle relation, certains textes privilégient telle construction plutôt que telle autre et, pour une construction donnée, tel texte va la spécialiser dans l'expression de telle relation ou non. On a eu trop tendance à confondre généralité de la relation et généralité des marqueurs. Or, ce que l'on recherche dans les textes, ce sont les marqueurs, et ces derniers sont des objets linguistiques, à ce titre intégrés dans un contexte (entendu ici dans un sens large, englobant tout ce qui va du texte au domaine lui-même) qui joue aussi son rôle.

Le choix de la formulation d'une relation est loin d'être stable et loin d'être neutre sur le plan énonciatif. Que dans nos corpus, le patron *est-un* manifeste sa meilleure précision – 40% – dans un manuel de parapente trouvé sur Internet n'est

sans doute pas un hasard. On peut émettre l'hypothèse que, destiné à un public large et ignorant tout du domaine, il a pour objectif premier d'en fixer les notions (mais ce n'est qu'une hypothèse parmi d'autres possibles, nous n'avons pas poussé suffisamment l'analyse pour étayer une explication satisfaisante).

Il serait en conséquence bon de rectifier la conception de la généralité et de se garder de l'appliquer aux marqueurs linguistiques qui sont recherchés dans les textes. Effectivement, certains permettront de construire des relations génériques et d'autres permettront de construire des relations spécifiques. Mais, dans le sens où ils ne sont pas neutres et entrent dans un contexte, dans le sens où ils sont pris dans des schémas, des habitudes langagières, ces marqueurs sont, eux, toujours spécifiques.

3. Par rapport à la mise au point d'un outil intégrant du TAL. Du point de vue de la construction de CAMÉLÉON, qui différencie une base globale et les bases particulières que l'utilisateur construit pour chaque projet, une conséquence immédiate de notre étude est que l'objectif de remplir la base globale de patrons aussi « tous-terrains » que possible n'a plus guère de sens. Compte tenu du coût de l'élaboration des patrons lexico-syntaxiques, il vaut mieux remplir cette base avec l'ensemble des patrons qui ont été élaborés, sans considération de leur performance, puisque notre étude montre que celle-ci peut varier du tout au tout, mais en les documentant, c'est-à-dire en les assortissant des diverses indications qui permettront à l'utilisateur de se faire une idée de la pertinence du patron pour son propre projet :

- le nom du patron, le jeu d'étiquettes avec lequel il a été élaboré (rappelons que CAMELEON peut accepter des textes étiquetés avec divers étiqueteurs, cependant il faut fournir le jeu d'étiquette en paramètre),
- la description du patron en langage intelligible,
- le patron lui-même,
- des exemples de contextes visés,
- des informations sur les corpus sur lesquels il a été testé,
- les résultats pour chaque corpus testé.

Les deux derniers points sont supposés aider l'utilisateur à décider d'employer ou non le patron pour son propre projet, mais l'avant-dernier n'est en fait pas si simple à documenter.

4. Par rapport aux genres textuels. L'idée sous-jacente est de donner à l'utilisateur des éléments pertinents de caractérisation des corpus utilisés pour l'élaboration des patrons, de telle sorte qu'il puisse apprécier la distance de son propre corpus avec chacun des corpus de test et qu'il puisse ainsi éventuellement extrapoler les résultats à ses propres textes. Toute la question tourne autour de la définition des éléments pertinents pour caractériser les textes. Comme elle occuperait facilement un article à elle seule, nous ne pouvons ici qu'esquisser quelques remarques.

La problématique du genre déborde largement le domaine du TAL, il s'agit de ranger chaque texte (instance) dans une classe plus globale (type) afin de généraliser au type certaines des observations menées sur l'instance ou d'affecter une instance quelconque à un type identifié. La définition du type / genre mobilise, selon le point de vue adopté, deux ensembles de critères : « la notion de genre est une notion biface qui fait correspondre une face interne (les fonctionnements linguistiques) avec une face externe (les pratiques socialement signifiantes) » (Branca-Rosoff, 1999 : 116). Schématiquement, on peut construire les genres en fonction de la relation des textes avec une pratique sociale (tradition rhétorique) ou en fonction de traits linguistiques internes aux textes (sur une opposition similaire, voir aussi Habert, 2000). C'est sur ce dernier point de vue que se fondent les travaux sur les typologies inductives (pour le français, voir Illouz *et al.*, 1999 ; Habert *et al.*, 2000 ; Folch *et al.*, 2000). Dans la perspective du TAL, l'objectif est de garantir une validité identique des traitements pour l'ensemble des textes rattachés aux types. On suppose ainsi que les types obtenus manifestent suffisamment de régularités au plan du fonctionnement linguistique pour faire des textes de chaque type un ensemble homogène et réagissant de façon identique à un traitement donné.

Or, quelle que soit la démarche adoptée (recours à l'induction ou référence à une pratique sociale), force est de constater qu'on est très loin d'un consensus – et c'est même un euphémisme – en matière de typologie textuelle. La définition même des critères devant entrer en jeu pour une telle typologie est un chantier en cours – par exemple (Malrieu et Rastier, 2001) vérifient sur corpus les correspondances entre un classement en genres et des agrégats de traits internes des textes tels que ceux pris en compte dans les travaux sur les typologies inductives – et on est à peine aux balbutiements d'une mise en relation de types avec des fonctionnements linguistiques qui permettrait de prédire un tant soit peu le comportement d'un texte non encore analysé par rapport à une tâche donnée. Le terrain commence à être exploré, comme nous l'avons signalé en introduction, mais un des principaux écueils tient à ce que les critères à prendre en compte fluctuent selon la tâche considérée : ce ne seront sans doute pas les mêmes indicateurs qui pourront rendre compte de différences de performances pour une tâche d'étiquetage morpho-syntaxique, de résolution d'anaphores, de rattachement prépositionnel ou de repérage de relations conceptuelles.

De plus, on est loin d'avoir la certitude qu'un même traitement produira des résultats identiques sur des textes qui se rangent dans le même genre[9]. Nos données tendraient même à montrer le contraire. Nous avons, par exemple, plusieurs manuels dans notre corpus mais nous n'observons pas avec eux d'homogénéité des scores de précision d'un patron donné. À l'heure actuelle, on peut en déduire soit qu'il n'y a pas de correspondance entre genre et fonctionnement linguistique, soit que le genre « manuel » est trop grossier et qu'il faudrait être plus précis c'est-à-dire déterminer

[9] Pour éviter toute confusion, rappelons que nous réservons le terme *genre* au classement relié à une pratique sociale et *type* au classement établi à partir de traits internes des textes.

des sous-genres, soit que le genre « manuel » n'a pas de sens et qu'il faut déterminer si ces textes présentent des caractéristiques linguistiques similaires, soit …

Une question essentielle à résoudre, à partir de ce genre d'observation, touche à l'identification des différents paramètres internes ou externes qui permettent d'expliquer cette disparité. Il pourrait être notamment intéressant de faire intervenir dans la recherche un troisième facteur, c'est-à-dire de croiser la façon dont les textes réagissent à un traitement donné avec leurs caractéristiques externes et internes. Par exemple, pour la tâche qui nous intéresse, constituer des regroupements des textes en fonction des taux de précision obtenus et voir ce que ces textes ont en commun.

Il est clair que l'amélioration des traitements automatiques passe par la définition de traitements plus différenciés et adaptatifs dont la mise au point implique des avancées en matière de typologie textuelle, et notamment la détermination des critères sur lesquels construire des types.

6. Conclusion

Nous avons rapporté ici un ensemble de résultats liés à la mise au point d'une base de marqueurs de relations lexicales pour un outil d'aide à la réalisation d'ontologies à partir de textes, CAMÉLÉON. L'évaluation de ces patrons sur huit corpus différents montre une grande variation de leurs performances selon le corpus testé. Cette étude a donc permis d'établir des résultats de trois ordres :

Du point de vue pratique, nous proposons une base documentée de 70 patrons lexico-syntaxiques évalués sur 8 corpus, avec, pour chaque patron, sa précision sur chacun des corpus de l'étude. Cette base, fournie en annexe, peut servir dans un contexte TAL plus large que celui de la construction de modèles de connaissances. De plus, elle est également disponible avec le logiciel CAMÉLÉON.

Du point de vue technique, l'analyse des résultats nous a conduites à revoir la notion de « base de marqueurs génériques » dans CAMÉLÉON, au profit d'une base de marqueurs réutilisables. L'idée a été mise en œuvre dans une nouvelle version du logiciel. Il est désormais possible de capitaliser chacune des expériences réalisées à l'aide de CAMÉLÉON, de conserver les marqueurs utilisés pour chaque projet avec leurs performances sur les corpus associés. Conformément à la proposition décrite en partie 5, alinéa 2, un marqueur n'est plus enregistré seul mais documenté par les différentes expériences au cours desquelles il a été utilisé.

Enfin, d'un point de vue méthodologique, cette étude contribue à une réflexion à double facette, tant pour la linguistique de corpus que pour le TAL. La variabilité souligne les enjeux et difficultés d'une approche de la linguistique par l'étude des corpus. Elle rappelle l'exigence de précision et prudence requise avant tout énoncé de résultats, et la nécessité de diversifier les études et les types de corpus. Pour le TAL, elle confirme qu'une meilleure qualité des traitements passe par leur différenciation et adaptation aux besoins et aux contextes. Dans les deux cas, un

besoin récurrent se fait ressentir : celui d'une meilleure caractérisation des textes, ou des genres textuels. Cependant, tout indique qu'une approche naïve qui irait vers une classification unique et rigide des textes, est vouée à l'échec. En effet, la diversité des écrits vient se combiner à celle des types d'analyses et de traitements effectués sur les textes.

Remerciements

Nous remercions Patrick Séguéla, concepteur de CAMÉLÉON, et les différents intervenants qui ont fait évoluer le logiciel depuis 2001 pour parvenir à sa version actuelle : Kévin Ottens, Cédric Delmas, Alexandre Simonnet. Sans eux, ce travail n'aurait pas pu être possible. Nous remercions aussi les différents relecteurs des premières versions de cet article qui nous ont permis de clarifier certains points.

7. Bibliographie

Aussenac-Gilles N., Seguela P., « Les relations sémantiques : du linguistique au formel ». *Cahiers de grammaire*, Numéro spécial sur la linguistique de corpus. A. Condamines (Ed.). Toulouse : Presse de l'UTM, Vol. 25., Déc. 2000, 175-198.

Biber D., *Variation accross speech and writing*, Cambridge, Cambridge University Press, 1988.

Borillo A., "Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie", *LINX*, Vol. 34-35, 1996, p. 113-124.

Branca-Rosoff S., "Des innovations et des fonctionnements de langue rapportés à des genres", *Langage et société*, Vol. 87, 1999, p. 115-129.

Condamines A., Rebeyrolle J., "Searching for and Identifying Conceptual Relasionships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results", in D. Bourigault, M.-C. L'Homme & C. Jacquemin (eds), *Recent Advances in Computational Terminology*, John Benjamins, 2000.

Condamines A, "Corpus Analysis and Conceptual Relation Patterns". *Terminology* 2002. p. 141-162.

Folch H., Heiden S., Habert B., Illouz G., Lafon P., Nioche J., Prévost S., "TypTex : Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation", In M. Gavrilidou et G. Carayannis (Eds.), *Language Resources and Evaluation Conference (LREC)*, Athènes, Grèce, 2000, p. 141-148.

Frérot C., Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambiguïtés de rattachement prépositionnel, Doctorat Nouveau Régime, Université Toulouse II Le Mirail, 2005.

Garcia, D., Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis. Thèse de doctorat en informatique, Université Paris IV, 1998.

Girju R., Moldovan D., "Text Mining for Causal Relations", *FLAIRS 2002*, Pensacola Beach, Florida, 14-16 mai 2002, p. 360-364.

Grabar N., Hamon T., "Les relations dans les terminologies structurées : de la théorie à la pratique", *Revue d'intelligence artificielle*, Vol. 18(1), 2004, p. 57-85.

Habert B., "Des corpus représentatifs : de quoi, pour quoi, comment ?" *Cahiers de l'Université de Perpignan*, Vol. 31, 2000, p. 11-58.

Habert B., Illouz G., Lafon P., Fleury S., Folch H., Heiden S., Prévost S., "Profilage de textes : cadre de travail et expérience", In M. Rajman (Ed.), *JADT (Journées Internationales d'Analyse Statistique des Données Textuelles)*, Lausanne, 2000.

Haddad M., Extraction et impact des connaissances sur les performances des systèmes de recherche d'information, Thèse de doctorat en Informatique de l'Université de Grenoble 1, 2002.

Hamon T. et A. Nazarenko A., "Detection of synonymy links between terms: experiment and results", *Recent Advances in Computational Terminology*. John Benjamins, 2001.

Hearst M., "Automatic Acquisition of Hyponyms from Large Text Corpora", *COLING-92*, Nantes, 23-28 août 1992, p. 539-545.

Illouz G., Habert B., Fleury S., Folch H., Heiden S., Lafon P., "Maîtriser les déluges de données hétérogènes", In A. Condamines, C. Fabre et M.-P. Péry-Woodley (Eds.), *TALN'99, Atelier thématique, Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, 12-17 juillet 1999, p. 37-46.

Jackiewicz A., "L'expression lexicale de la relation d'ingrédience (partie-tout)", *Faits de langues*, Vol. 7, 1996, p. 53-62.

Jacques M.-P., Soubeille A.-M., "Partages des termes, partage des connaissances ? Construire une modélisation unique de plusieurs corpus", *IC'2000 Journées Francophones d'Ingénierie de la connaissance*, Toulouse, 10-12 mai 2000, Institut de Recherche en Informatique de Toulouse, p. 313-324.

Maedche A., Staab S., "Mining Ontologies from Text", *EKAW 2000*, Juan-les-Pins, France, 2000, Springer, p. 189-202.

Malrieu D., Rastier F., "Genres et variations morphosyntaxiques", *Traitement Automatique des Langues*, Vol. 42(2), 2001, p. 547-577.

Morin E., "Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique", *Traitement Automatique des Langues*, Vol. 40(1), 1999, p. 143-166.

Mothe J., Tanguy L., "Linguistic features to predict query difficulty - a case study on previous TREC campaigns", *ACM - SIGIR 2005, Predicting Query Difficulty - Methods and Applications Workshop*, Salvador de Bahia, Brésil, 15-19 août 2005, ACM, p. 7-10.

Rebeyrolle J., Tanguy L., "Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires", *Cahiers de Grammaire*, Vol. 25, 2001, p. 153-174.

Riloff E., "Automatically Generating Extraction Patterns from Untagged Texts", *Thirteenth National Conference on Artificial Intelligence (AAAI'96)*, 1996, p. 1044-1049.

Séguéla P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, Thèse de doctorat en Informatique de l'Université Toulouse III, mars 2001.

Tchalakova M., Popov B., Yankova M., "Methodology for Bootstrapping Relation Extraction for the Semantic Web", *AIMSA 2006*, Varna, Bulgaria, 13-15 septembre 2006 à paraître.

Yamaguchi T., "Acquiring Conceptual Relationships from Domain-Specific Texts", *17th International Joint Conference on Artificial Intelligence (IJCAI'2001) - Ontology Learning Workshop*, Seattle, USA, 4 août 2001, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-38/Yamaguchi_11-submission.pdf.

Annexe : les patrons lexico-syntaxiques

Les étiquettes morpho-syntaxiques sont celles de la version courante du TreeTagger. Chaque élément du patron est séparé par une espace. L'expression des patrons suit la syntaxe des expressions régulières plus d'autres particularités : le symbole * seul signifie un joker indéfini (on ne précise pas le nombre de mots) ; le symbole > avant une forme signifie que la forme est celle du lemme ; lorsqu'une forme est associée à une série d'étiquette (pas d'espace entre la forme et les étiquettes), cela signifie que la forme doit être de la catégorie stipulée par l'une des étiquettes, par exemple, le patron *dire* correspond au lemme *dire* ou *dit* de catégorie adjectif ou verbe au futur, à l'imparfait, au passé simple ou au présent ou participe passé. La première ligne indique le nom du patron et la fourchette de variation de son taux de précision (en pourcentage) sur les 8 corpus étudiés.

Définition

definir 98-100
>^(définir|défini)$ {SYM|NOM|NAM|ABR|ADV|DET.*|PRO.*|INT|NUM|ADJ|PRP.*|KON}* ">^comme$" PRO:POS|NAM|SYM|NOM|DET.*|ADJ|INT|PRO:DEM|PRO:REL|PRO:IND|ABR|NUM|VER.*|KON

denom1 58-96
>^(appeler|appelé|nommé|nommer|désigner|désigné|dénoter|dénoté|dénommer|dénommé|baptiser|baptisé)$

denom2 10-100
>^(porter|appliquer|employer|réserver|recevoir|prendre|utiliser|donner|proposer|mériter)$ 6 ">^[cl]e$" ">^(nom|terme|mot|expression|vocable|appellation|désignation|dénomination)$"

denom3 0-100
>^être$ 1 ">^[cl]e$" ">^nom$"

denom4 0-100
>^[cl]e$ ">^(nom|terme|mot|expression|vocable|appellation|désignation|dénomination)$" * ">^(donné|donner|porté|porter|appliqué|appliquer|employé|employer|réservé|ré

server|reçu|recevoir|pris|prendre|utilisé|utiliser|proposé|proposer|recouvrir|regrouper|grouper)$"

denom5 (pas d'occurrences)
>^avoir$ ">^pour$" ">^nom$"

denom6 67-100
>^(sous|où)$ ">^[cl]e$" ">^(nom|terme)$"

denom7 0-100
>^(parler|qualifier)$ {PRO.*|NOM|INT|KON|DET.*|ADV|ADJ|NUM|NAM|ABR|PRP.*|SYM}* ">^d[eu]$"

denom8 25-100
>^(nom|terme|mot|expression|vocable|appellation|désignation|dénomination|concept|notion|idée)$ * ">^(référer|renvoyer|référence|allusion)$" {ADV}1 ">^à$"

dire 0-82
ADJ|VER:futu|VER:pper|VER:simp|VER:impf|VER:pres">^(dire|dit)$"

entendre_par 33-100
>^entendre$ 6 ">^par$"

etre1 17-23
PRO:REL|PRO:IND|INT|PRP.*|ABR|NOM|VER.*|NUM|KON|PRO:POS|SYM|NAM|DET.*|ADJ|ADV VER:futu|VER:ppre|VER:pres|VER:impf|VER:simp">^être$" {KON|ADV}1 DET.*|PRO:DEM|NUM|PRP.*|

etre2 0-11
PRP.*|NOM|SYM|DET.*|INT|PRO:IND|ABR|ADV|PRO:REL|ADJ|NUM|KON|NAM|PRO:POS|VER.* VER:simp|VER:impf|VER:futu|VER:ppre|VER:pres">^être$" ABR|NAM|PRO:IND|PRO:POS|NOM|SYM|VER:ppre|INT 6 PRO:DEM|NUM|DET.*

etre3 4-26
>^ce$ VER:impf|VER:futu|VER:ppre|VER:pres|VER:simp">^être$" {KON|ADV}1 NUM|DET.*

il s'agit 3-13
>^il$ {">^se$"}1 "agit$"

par_*_entendre 75-100
>^par$ 5 ">^entendre$"

signifier 14-38
ADJ|ADV|VER.*|PRO:REL|ABR|NOM|NAM|NUM|KON|INT|PRP.*|DET.*|SYM ">^signifier$" ">^[^q]"

vouloir_dire 0-100
>^vouloir$ ">^dire$"

Hyperonymie

distinguer_plusieurs_X_tel_que_Y (pas d'occurrences)
VER:ppre|VER:impf|VER:simp|VER:pres|VER:futu">^(distinguer|reconnaître|différencier|discriminer|isoler|séparer|discerner|remarquer)$" 4 ">^plusieurs$" {DET.*|NOM|KON|ADV|PRP.*|ADJ}* ">^(tel|dont)$"

divers_X_comme_Y 100
>^divers$ {ADV|PRP.*|NOM|DET.*|ADJ|KON}* ">^comme$"

et_Adv_Y 0-58
>^et$ 1 ADV">^(notamment|notablement|spécialement|particulièrement|surtout)$"

et_Adv_Y2 0-100
>^et$ 1 ADV">^(notamment|notablement|particulièrement|spécialement)$"

et_Adv_Y3 0-25
>^et$ 1 ">^(par|en|avant)$" ">^(exemple|particulier|tout)$"

et_autres 0-71
>^(et|ou)$ {">^(de|tout)$"}1 ">^autre$"

etre_le_plus_de_tous_les_X 0
VER:ppre|VER:pres|VER:impf|VER:simp|VER:futu">^être$" 1 DET:ART">^le$" ">^(plus|moins)$" 1 ">^de$" ">^tout$" ">^le$"

etre_le_X_le_plus 0-86
VER:simp|VER:ppre|VER:impf|VER:futu|VER:pres">^être$" 1 DET:ART">^le$" {PRP.*|NOM|ADV|ADJ|KON|DET.*}* DET.*">^le$" ">^(plus|moins)$"

etre_un 11-40
ADJ|SYM|DET.*|NAM|KON|NOM|ABR|VER.*|ADV|PRP.*|PRO:REL|INT|NUM|PRO:IND|PRO:POS VER:futu|VER:ppre|VER:pres|VER:simp|VER:impf">^être$" {ADV|KON}1 PRP:det|PRO:DEM|DET.*|NUM

etre_un_X_très 0-67
VER:impf|VER:futu|VER:simp|VER:pres|VER:ppre|VER:pper">^être$" 1 DET:ART {ADJ|ADV|PRP.*|NOM|KON}6 ">^(très|absolument)$"

etre_un_X_très2 0-29
VER:ppre|VER:futu|VER:pres|VER:pper|VER:impf|VER:simp">^être$" 1 DET:ART * ">^(très|absolument)$"

inclusion1_X_,_Y 100
>^parmi$ {">^tout$"}1 DET:ART|PRP:det {PRP.*|NOM|ADJ|DET.*|KON|ADV}* ">^,$" DET:ART {PRP.*|ADJ|NOM|ADV|DET.*|KON}* VER:subi|VER:pres|VER:ppre|VER:futu|VER:simp|VER:cond|VER:subp|VER:impf

inclusion2_X_,_Y (pas d'occurrences)
^(dans|au)$ {DET:ART}1 "^(ensemble|nombre)$" PRP:det {DET.*|KON|PRP.*|ADV|ADJ|NOM}* ">^,$" DET:ART {PRP.*|ADV|NOM|ADJ|KON|DET.*}* VER:cond|VER:futu|VER:pres|VER:ppre|VER:subp|VER:subi|VER:simp|VER:impf

le_plus_adj_des_X_,_soit_Y (pas d'occurrences)
DET.*">^le$" ">^(plus|moins)$" ADJ PRP.*">^(de|du)$" {ADJ|NOM|ADV|PRP.*|KON|DET.*}* PUN.*">^,$" ">^(c'est-à-dire|soit)$"

sorte_de 33-100
VER:simp|VER:pper|VER:impf|VER:futu|VER:pres|VER:ppre">^être$" 1 DET:ART ">^(sorte|type|genre|style|variété|espèce)$" ">^de$"

tout_autre_type_de 0
>^tout$ ">^autre$" ">^(type|genre|sorte|espèce|variété|style)$" ">^de$"

tout_X_est_un_Y (pas d'occurrences)
>^tout$ NOM|ADJ {DET.*|PRP.*|ADV|NOM|KON|ADJ}* VER:pres|VER:impf|VER:ppre|VER:futu|VER:simp">^être$" ">^(un|du)$" DET.*|PRP:det

utiliser_Y_comme_X 55-100
VER:pres|VER:ppre|VER:simp|VER:futu|VER:impf">^utiliser$" {NOM|ADJ|ADV|KON|DET.*|PRP.*}* ">^en$" ">^tant$" ">^que$"

utiliser_Y_en_tant_que_X(pas d'occurrences)
VER:impf|VER:futu|VER:ppre|VER:simp|VER:pres">^utiliser$" {ADJ|ADV|DET.*|KON|NOM|PRP.*}* ">^en$" ">^tant$" ">^que$"

X_,_adv_exc1_Y 0-67
ADJ|NOM PUN.*">^,$" "^(sauf|hormis|excepté)$" PRO:DEM|DET.*|PRP:det

X_,_adv_exc2_Y 0-42
NOM|ADJ PUN.*">^,$" 2 "^exception$" 2 PRO:DEM|DET.*|PRP:det

X_,_Adv_Y 0-100
NOM|ADJ PUN.*">^,$" ADV">^(notamment|notablement|spécialement|particulièrement|surtout)$" PRP:det|DET.*

X_,_Adv_Y2 20-100
NOM|ADJ PUN.*">^,$" ">^(par|en|avant)$" ">^(exemple|particulier|tout)$" DET.*|PRP:det

X_,_tout_en_comptant_Y (pas d'occurrences)
ADJ|NOM PUN.*">^,$" ">^tout$" ">^en$" ">^comptant$"

X_,_Y_Adv 0-50
ADJ|NOM PUN.*">^,$" DET.*|PRP:det {PRP.*|ADV|NOM|KON|ADJ|DET.*}* ADV">^(notamment|notablement|spécialement|particulièrement|surtout)$"

X_,_Y_Adv2 0-100
ADJ|NOM PUN.*">^,$" DET.*|PRP:det {PRP.*|NOM|ADV|ADJ|DET.*|KON}* ">^(par|en|avant)$" ">^(exemple|particulier|tout)$"

X_,_y_compris_Y 0-100
NOM|ADJ PUN.*">^,$" ">^y$" "^compris$" PRO:DEM|PRP:det|DET.*

X_?_desquels_Y (pas d'occurrences)
ADJ|NOM "^au$" {">^premier$"}1 ">^(nombre|rang)$" ">quel$"

X_est_special_de_Y 0-50
NOM|ADJ VER:ppre|VER:simp|VER:pres|VER:impf|VER:futu">^être$" 4 ">^(spéc
ification|spécialisation|précision|détermination)$" ">^(de|du)$"

X_parmi_lesquels_Y 0-100
ADJ|NOM ">^parmi$" ">^lequel$" DET.*|PRP.*

X_se_présente_sous_forme_de_Y 67-100
NOM|ADJ ">^se$" VER:simp|VER:impf|VER:pres|NOM|VER:futu|ADJ|VER:ppre"
>^présenter$" 2 ">^(sous|en)$" 3 ">^forme$"

X_v_special_de_Y 0
NOM|ADJ VER:ppre|VER:pres|VER:simp|VER:impf|VER:futu">^(former|représent
er|constituer|consister)$" 4 ">^(spécification|spécialisation|précision|détermination)$
" ">^(de|du)$"

Y_app_classe_X 0-100
ADJ|NOM VER:pres|VER:futu|VER:impf|VER:simp|VER:ppre">^(appartenir|resse
mbler|dériver)$" 2 ">^(à|de|du)$" 1 ">(classe|caste|catégorie|groupe|division|espèce|s
orte|race|ensemble|variété|type|modèle|famille|genre|collection|partie)"

Y_virgule_le_plus_adj_des_X 0
ADJ|NOM PUN.*">^,$" DET:ART">^le$" ">^(plus|moins)$" ADJ PRP.*">^(de|du
)$" PRP.*|NOM|ADJ|KON|ADV|DET.*

Y_virgule_le_X_le_plus 0-100
NOM|ADJ PUN.*">^,$" DET:ART">^le$" {NOM|ADJ|ADV|DET.*|PRP.*|KON}*
DET:ART">^le$" ">^(plus|moins)$" {KON|PRP.*|NOM|DET.*|ADV|ADJ}* PUN.
*">^,$"

Méronymie

composition_parties_identiques 43-100
VER:pper">^(réunir|rassembler|unir|regrouper|grouper|collecter)$" 2 ">^(en|dans|sur
|relatif)$"

N_compo_non_org 67-75
>^(tas|amas|ramassis|masse|accumulation|entassement)$ 1 ">^(de|du)$" {PRP.*|NO
M|KON|DET.*|ADV|ADJ}* ">^dans$"

npartie_de 0-100
VER:pres|VER:ppre|VER:simp|VER:futu|VER:impf">^être$" {DET.*|ADV}2 ">^(c
onstituant|composant|composante|ingrédient|membre|organe|élément|partie|bout|case|
division|fraction|fragment|morceau|parcelle|part|pièce|portion|étape|période|phase|stad
e)$" "^(de|des)$"

X_compo_identique 0-77

VER:impf|VER:simp|VER:futu|VER:ppre|VER:pres">^(réunir|unir|rassembler|regrouper|grouper|collecter)$" 2 DET:ART|PRP:det

X_etre_decompose_en_Y 60-100
VER:simp|VER:impf|VER:pres|VER:futu|VER:ppre">^être$" 1 VER:pper">^(analyser|décomposer|démembrer|désassembler|disjoindre|dissocier|désagréger|séparer|découper|couper|partager|trancher|diviser|fractionner|fragmenter)$" 2 ">^en$"

X_etre_nom_groupe_de_Y 50-100
VER:pres|VER:futu|VER:impf|VER:simp|VER:ppre">^être$" {ADV}2 DET:ART|PRP:det ">^(ensemble|groupe|classe|réunion|famille|collection)$" 2 ">^(de|du)$"

X_formé_de_Y 17-91
VER:pper">^(former|constituer)$" 2 ">^(de|du)$"

X_inclure_Y 19-100
NOM|ADJ VER:simp|VER:futu|VER:pres|VER:impf|VER:ppre">^(abriter|comporter|comprendre|compter|inclure|intégrer)$"

X_renferme_Y 27-94
NOM|ADJ NOM|VER:futu|VER:ppre|ADJ|VER:impf|VER:pres|VER:simp">^(renfermer|contenir|englober|abriter|emprisonner|loger|incorporer)$"

X_se_decomp_en_Y 0-100
NOM|ADJ ">^se$" VER:pres|VER:impf|VER:ppre|VER:futu|VER:simp">^(analyser|décomposer|démembrer|désassembler|disjoindre|dissocier|désagréger|séparer|découper|partager|trancher|diviser|fractionner|fragmenter|dissoudre)$" 3 ">^en$"

Y_etre_classe_dans_X 33-100
VER:pres|VER:futu|VER:simp|VER:ppre|VER:impf">^être$" VER:pper">^(classer|classifier|cataloguer|ranger|placer|inclure|étiqueter|catégoriser|grouper)$" 2 ">^(en|dans|intérieur|parmi|coeur|centre|milieu|dedans|fond|sein|sur)$"

Y_etre_interne_à_X 0
VER:pres|VER:impf|VER:simp|VER:futu|VER:ppre">^être$" {PRP.*|DET.*|ADV}3 ">^(interne|intérieur)$" ">^à$"

Y_rapprocher_dans_X 0
VER:pper">^(rapprocher|recueillir)$" 2 ">^(dans|sur|en|vers)$"

Y_verbe_constitution_classe_X0-67
VER:pres|VER:ppre|VER:simp|VER:futu|VER:impf">^(former|représenter|constituer|consister)$" 2 ">(classe|caste|catégorie|groupe|division|espèce|sorte|race|ensemble|variété|type|modèle|pattern|famille|genre|collection|concept|partie|partition)"

Reformulation

X_cad_Y 29-100
ADJ|NOM {">^(à|,)$"}2 "^(c'est-à-dire|savoir)$" PRP:det|DET.*

Varia

baser 75-100
VER:pper">^(fonder|baser)$" ">^sur$"

Y_situé_dans_X 0-75
VER:pper">^(situer|localiser|placer|positionner|disposer|insérer|installer|poser|accrocher|fixer|sceller|attacher|mettre|caser|nicher|déposer|ficher|loger|adosser|appliquer|camper|jucher|flanquer|arrimer|amarrer|boulonner|clouer|coincer|enchâsser|épingler|river|riveter|visser|implanter|ranger|garer|immobiliser|introduire|enfoncer|plonger|enfouir|ensevelir|contenir|emprisonner|abriter|dresser|étendre|trouver)$" 3 ">^dans$"