La syntaxe comme marche-pied de l'acquisition de connaissances : bilan critique d'une expérience*

Benoît Habert et Adeline Nazarenko
ELI-ENS de Fontenay-St Cloud
31 av. Lombart F-92260 Fontenay-aux-Roses `Nom.Prénom@ens-fcl.fr`

Thèmes Extraction de connaissances à partir de corpus : utilisation de méthodes linguistiques, évaluation des méthodes d'acquisition de connaissances

1 Introduction

Belle postérité que celle de Harris (1904-1992). Sa conception "distributionnaliste" de la sémantique – le sens des mots se déduit des constructions dans lesquelles ils figurent – a fait florès. Sous une version affaiblie dans les travaux statistiques sur les collocations ([Church and Hanks, 1990] en est le meilleur exemple). Ou sous une version plus structurelle [Hindle, 1990, Ruge, 1991, Grefenstette, 1994, Grishman and Sterling, 1992]. Cette hypothèse semble enfin "testable" à vaste échelle. Les possibilités de stockage et de calcul actuelles autorisent les essais de la version affaiblie, et les progrès de l'analyse syntaxique "robuste" (partielle, probabiliste etc.) ceux de la version structurelle.

Pour autant, peut-on s'attendre à produire des résultats équivalents à ceux de Harris ou de l'équipe de N. Sager sur le discours médical ? Cette approche

*Les travaux du groupe Terminologie et Intelligence Artificielle ont contribué à la mise au point de l'approche présentée. Nous remercions en outre pour leurs précieuses et précises remarques sur des états de ce travail Christian Jacquemin (IRIN-Nantes), Fernande Dupuis et Philippe Barbaud (UQAM), Cécile Fabre (IRISA), Serge Fleury et Elie Naulleau (ELI - ENS de Fontenay St Cloud), Houssem Assadi (DER-EDF) et Anne Condamines et Marie-Paule Pery-Woodley (ERSS - Université de Toulouse le Mirail). Roland Dachelet nous a aidés à mieux comprendre l'optique des travaux de l'équipe de N. Sager. GraphX, développé par Serge Heiden, nous a permis de visualiser et de manipuler nos graphes. Nous remercions particulièrement Didier Bourigault (DER-EDF), à la fois pour nous avoir fourni les résultats de LEXTER sur le corpus étudié et pour avoir examiné et discuté les données issues de la simplification d'arbres et les analyses en découlant.

débouche, selon ses partisans, sur la mise en évidence de classes d'opérateurs et de classes d'arguments, qui sont supposées fortement liées aux notions et relations du domaine considéré. Serait-ce alors une voie presque "royale" pour l'acquisition de connaissances à partir de corpus, voire pour la construction d'ontologies ?

Nous avons utilisé une approche très similaire à celle prônée par la tradition harrissienne sur un corpus concernant la médecine coronarienne, donc proche de ceux travaillés par Harris ou par N. Sager. Nous simplifions et rapprochons les groupes nominaux fournis par un analyseur syntaxique robuste (2ème partie). L'examen critique des résultats obtenus (3ème partie) conduit à la prudence, et, en tout cas, à caractériser plus précisément les conditions d'une induction de classes et relations sémantiques guidée par la syntaxe ainsi que les retombées que l'on peut en escompter (4ème partie).

2 Méthode : simplification de groupes nominaux

2.1 Obtenir les classes d'opérateurs et d'arguments à partir de phrases élémentaires

Dans la lignée des travaux de Harris, l'équipe de Naomi Sager a développé une méthodologie de mise au jour des classes sémantiques d'un sous-langage lié à un domaine d'activité donné à partir d'une analyse distributionnelle en phrases élémentaires [Sager *et al.*, 1987]. La première étape consiste à analyser manuellement un corpus restreint du sous-langage choisi. Chaque phrase se voit associer un arbre d'analyse. Cet arbre est ensuite simplifié pour restituer les anaphores, éliminer les coordinations, les enchâssements etc. Des transformations sont appliquées pour normaliser les phrases (passage du passif à l'actif etc.). Le résultat est un ensemble de phrases élémentaires. L'examen de ces phrases élémentaires fournit les opérateurs de base du sous-langage, et pour chaque opérateur, les classes de mots qui peuvent être en position d'arguments. Ces ensembles s'avèrent être autant de classes sémantiques. Cette démarche a été appliquée au discours médical [Sager *et al.*, 1987] et au discours pharmacologique.

2.2 Obtention d'arbres nominaux élémentaires

Notre recherche de classes sémantiques repose sur la simplification d'arbres syntaxiques fournis par un analyseur robuste, ici LEXTER (Bourigault 93). LEXTER extrait d'un corpus les groupes nominaux qui peuvent constituer des termes. Il divise chacune de ces séquences nominales en T(ête) et E(xpansion), de façon récursive. La séquence *stenose severe de le tronc commun gauche* est par exemple analysée comme suit : [T [T stenose][E severe]][E de le [T [T tronc][E commun]][E gauche]] [1]. Quand une séquence peut être découpée de plusieurs manières, LEXTER recherche dans le reste du corpus des situations non ambiguës permettant de choisir entre les analyses en compétition.

Pour mettre en évidence les comportements syntaxiques majeurs de chaque entrée lexicale et dégager certaines des relations sémantiques que ces entrées entretiennent, un programme, SYCLADE [2], s'appuie sur les arbres dérivés eux-mêmes pour en extraire les arbres élémentaires qui y figurent [3].

Deux types de simplifications peuvent être appliquées à un arbre dérivé donné :

- "Dé-concaténation" : chacun des sous-arbres dominés par la racine de l'arbre examiné est considéré à tour de rôle. Cela revient à "défaire" l'action de la règle qui a rassemblé ces arbres en un seul constituant.

- Elimination des modifieurs ou "désadjonction" [4].(cf. infra) : dans l'arbre entier, tous les constituants qui modifient un autre constituant sont éliminés.

Cette simplification s'arrête quand on considère que l'arbre examiné est *élémentaire*. Cela signifie qu'au début du traitement, on déclare quels sont les arbres qui ne doivent pas être simplifiés davantage. Dans la perspective qui est la nôtre, il s'agit d'arbres mettant en évidence une relation binaire entre deux mots "pleins", par exemple [SN [SN N][SP Prep [SN N]]] (*taux de mortalite*), [SN [SN N][SAdj Adj]] (*mort subite*), [SN [SAdj Adj][SN N]] (*faible pression*).

L'arbre fourni par LEXTER est donc transformé en [SN [SN [N stenose] [SAdj [Adj severe]]][SP [Prep de] [SN [Det le] [SN [SN [N tronc][SAdj [Adj commun]]] [SAdj gauche]]]]] par des méta-règles [Jacquemin, 1994]. La simplification selon ces principes de *stenose serre de le tronc commun gauche* produit quatre arbres élémentaires : *stenose severe*, *stenose de tronc*, *tronc commun*, *tronc gauche*.

2.3 Situation effective par rapport à l'approche harrissienne

Nous avons utilisé la même approche. Nous avons travaillé sur un corpus de langage spécialisé. Mais nous nous en sommes éloignés sur deux points. D'abord, nous nous sommes limités à l'analyse de groupes nominaux dont l'interprétation

1. Les accents sont enlevés par le programme d'étiquetage utilisé, qui substitue aux formes leur lemme. Les exemples fournis sont par conséquent lemmatisés (à l'exception des déterminants) et désaccentués. Nous avons en outre remplacé les mots contractés comme *du* par la suite préposition déterminant.

2. Développé par B. Habert

3. Cf. [Habert *et al.*, 1996] pour une présentation détaillée de l'algorithme.

4. Cet algorithme suppose de représenter un modifieur comme un (sous-)arbre dont un des fils porte la même étiquette que la racine de l'arbre et qui comporte plusieurs fils. Oter le modifieur consiste alors à remplacer la racine de l'arbre par le constituant dominé de même catégorie A l'inverse de l'adjonction des grammaires d'adjonction [Abeillé, 1993], d'où le nom de "désadjonction"

est plus délicate que celle des phrases verbales. En second lieu, notre effort a porté sur la mise au point d'une chaîne automatique de traitement, à la place de l'analyse manuelle opérée par N. Sager et ses collaborateurs, analyse manuelle limitée qui restreint l'utilisation de cette méthodologie pour de nouveaux domaines. Nous avons restreint la "normalisation" des arbres produits par l'analyseur syntaxique aux deux types de simplification présentés supra (de-concaténation, désadjonction), sans utiliser toutes les transformations possibles (pour ramener par exemple *severite de lesion* à *lesion severe*).

2.4 Caractérisation des entrées lexicales et des relations entre elles

Les collocations que mettent en évidence les arbres élémentaires correspondent, sauf erreur dans l'analyse initiale, à des relations syntaxiques avérées, à la différence d'une partie des travaux sur les collocations qui examinent les rapprochements à l'intérieur d'une "fenêtre" de quelques mots, sans étiquetage [Church and Hanks, 1990] ou avec étiquetage préalable [Daille, 1994] ou postérieur [Smadja, 1993]. Ici, dans *stenose severe de le tronc gauche*, *gauche* ne sera pas mis en relation avec *stenose*, puisqu'il ne modifie pas ce nom.

Les arbres élémentaires permettent de constituer des classes de contextes syntaxiques, prolégomènes à des catégories sémantiques, voire conceptuelles. Ainsi, à partir de l'arbre correspondant à *stenose severe*, on peut créer deux classes de contextes. La première, [SN [N stenose][SAdj [Adj ~]]] [5] comprend l'entrée *severe*, la seconde, [SN [N ~][SAdj [Adj severe]]], l'entrée *stenose*. Au fur et à mesure de l'examen des arbres élémentaires dégagés par l'algorithme de simplification, de nouvelles classes sont créées et les classes existantes complétées.

Si l'on considère les formes comme les noeuds d'un graphe non orienté, une construction partagée peut représenter une arête reliant deux noeuds. Par exemple, ~ *residuel* relie *lesion* à *stenose* (tous deux peuvent y occuper la position du nom, figurée par ~), et relie aussi chacun d'eux à *angor* et *image*. On calcule l'ensemble des arêtes entre les noeuds du graphe (ici les 1051 entrées lexicales figurant dans les 1252 groupes nominaux fournis par LEXTER). On extrait ensuite les sous-parties connexes de ce graphe [6].

5. Le tilde sert, comme dans les dictionnaires bilingues, pour remplacer l'entrée dans l'exemple. Ici il remplace *severe*.

6. Tels qu'il existe toujours une chaîne (une suite d'arêtes) reliant deux noeuds quelconques.

3 Induction de notions et de relations en médecine coronarienne

3.1 Le corpus MENELAS

Nous avons testé notre démarche sur un corpus utilisé par le Groupe de Travail Terminologie et Intelligence Artificielle (PRC-GDR Intelligence Artificielle, CNRS) pour comparer différentes méthodes d'extraction terminologique et de construction de bases conceptuelles et terminologiques. Ce corpus a été constitué pour le projet MENELAS de compréhension de dossiers médicaux [Zweigenbaum, 1994]. Il comprend 20 840 occurrences. Il a pour unité thématique les maladies coronariennes.

3.2 Repérage et caractérisation de notions

Les six composantes connexes produites par SYCLADE[7] sur ce corpus filtrent un certain nombre d'entrées lexicales qui correspondent pour partie à des notions fondamentales du domaine traité. Le premier sous-graphe connexe, décomposé en deux figures pour plus de lisibilité, correspond à 37 noms et 49 arêtes.

Dans la figure 1, les liens manifestent l'existence d'une famille de formes au fonctionnement proche : *plaque occlusion stenose lesion calcification spasme obstruction atteinte*, que nous appellons les `/affections corporelles localisées/`. La relation de localisation y est généralement présente, sous trois modalités : *de (artere tronc)*, *de N* où *N* est un nom d'artère (*carotide, interventriculaire)*, adjectif lié à une artère spécifique (*coronaire, coronarien, diagonal, circonflexe*). Le degré est un attribut fondamental : *significatif non-significatif severe important severite*. Ces affections renvoient, semble-t-il, à la fois à des `processus` (sauf *plaque*) : *persistance de (occlusion stenose)*, *(stenose lesion) residuel*, et à des entités concrètes et mesurables : *existence de (calcification lesion)*, *diametre de obstruction*.

Le reste de la première composante connexe (figure 2) complète la vision des sites esquissée. *Carotide, interventriculaire* et *branche* rejoignent *artere* dans les entités visées par des affections : *spasme sur (artere branche)*, *stenose de (artere carotide interventriculaire)*.

Un deuxième groupe de mots *infarctus myocarde necrose territoire ischemie* se caractérise par un mode de localisation propre *anterieur inferieur apical lateral* et les attributs d'un processus situé dans le temps : *absence de (ischemie infarctus)*, *sequelle de infarctus necrose)*, *infarctus (recent ancien)*.

7. Il génère les noeuds et les arêtes pour le logiciel GraphX de manipulation interactive de graphes sous X Window développé par Serge Heiden (`slh@ens-fcl.fr`), mis dans le domaine public.

Enfin, un petit ensemble d'`/interventions/` se dessine : *angioplastie revascularisation pontage* : *indication de* ~, *possibilite de* ~.

3.3 Repérage et caractérisation de relations

Dans la première composante connexe, les constructions manifestent en outre la relation `/a pour partie/` entre *artere* et *tronc* ou *segment* : *(segment tronc) de artere.*

Le réseau des adjectifs (figure 3), renseigne plutôt, lui, sur les relations et attributs à attacher aux notions. Il est dominé par les localisants, en particulier ceux qui renvoient à des artères : *marginal coronaire circonflexe diagonal coronarien.* L'opposition *droit gauche* constitue en fait une localisation auxiliaire, dans des contextes comme *artere circonflexe gauche.* Autre localisation, distincte, celle qui concerne la famille d'*infarctus* : *anterieur apical inferieur.* Si *artère* comme *infarctus* entrent dans une relation de `localisation`, celle-ci se réalise donc de deux façons différentes.

Le réseau permet également de préciser les conditions d'emploi de *coronaire* et *coronarien.* Malgré le nombre élevé (7) de contextes partagés et leur appartenance commune à la clique[8] *coronaire coronarien diagonal circonflexe*, les deux adjectifs ne sont pas synonymes : *coronarien* est associé à des adjectifs évaluatifs *severe significatif important*, ce qui n'est pas le cas de *coronaire.*

Ce réseau comprend enfin des antonymes : *residuel severe* et *important minime* qui peuvent exprimer la relation de `degré` à attacher à certaines affections.

4 Bilan critique

4.1 Atouts d'une approche endogène

La méthode de proposition de relations sémantique proposée est totalement endogène : aucune connaissance extérieure au corpus étudiée n'est exploitée[9].

Néanmoins, les relations sémantiques mises en évidence ne sont pas généralisables. Les oppositions manifestées sont propres au corpus examiné. Rien ne permet d'extrapoler des propriétés en langue à partir du fonctionnement des unités dans ce discours spécifique[10]. De la même manière, les classifications et ontologies élaborées à partir de ces résultats ne sont pertinentes que pour le corpus étudié.

L'ancrage discursif de nos résultats n'est cependant pas un défaut mais plutôt un atout. Même dans un domaine scientifique comme la pharmacie ou la médecine, les notions ne sont pas toujours stabilisées et le sens des concepts évolue

8. Une clique est un graphe dans lequel chaque noeud est relié par une arête à tous les autres.
9. Dans la continuité des choix de LEXTER.
10. Pour les corpus de taille modeste sur lesquels nous travaillons, on ne peut pas faire d'extrapolation statistique.

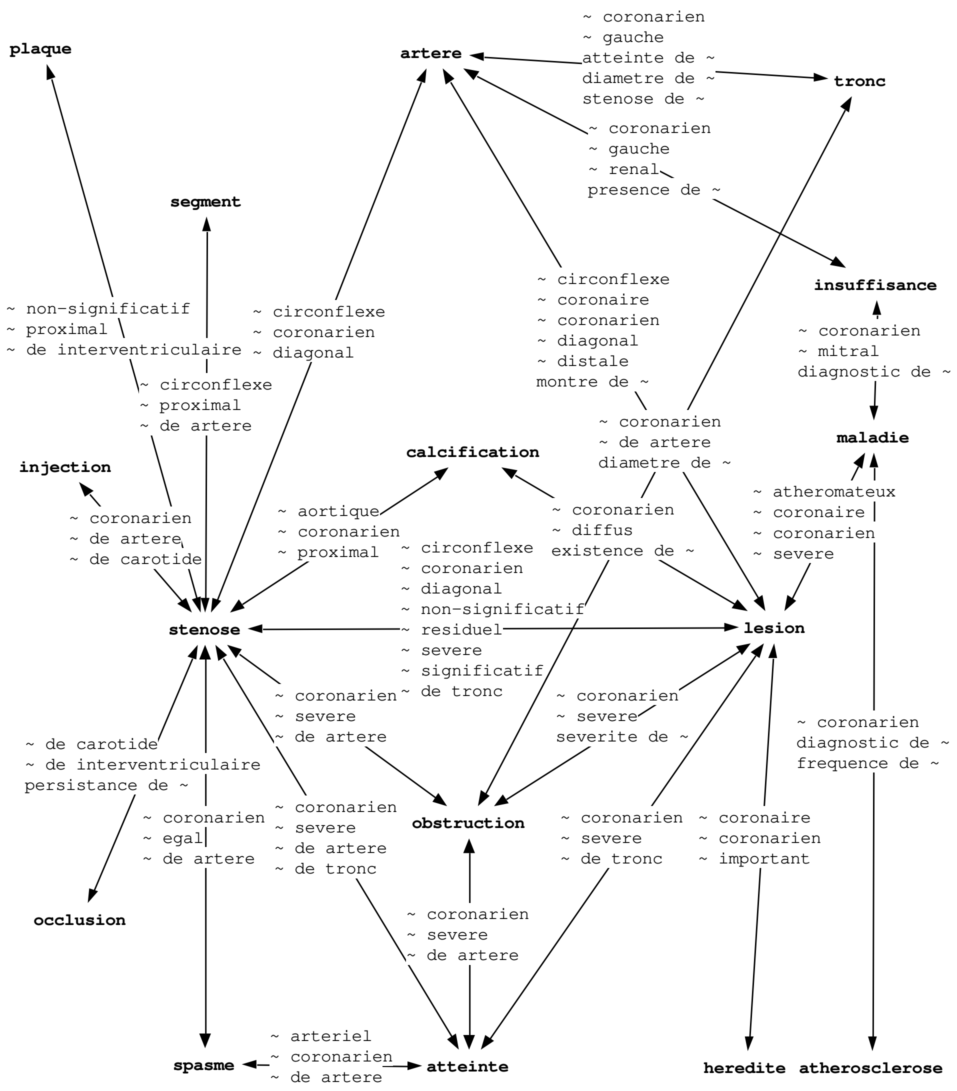

FIG. 1 - *1ère composante connexe (début), les* `/affections corporelles localisées/`

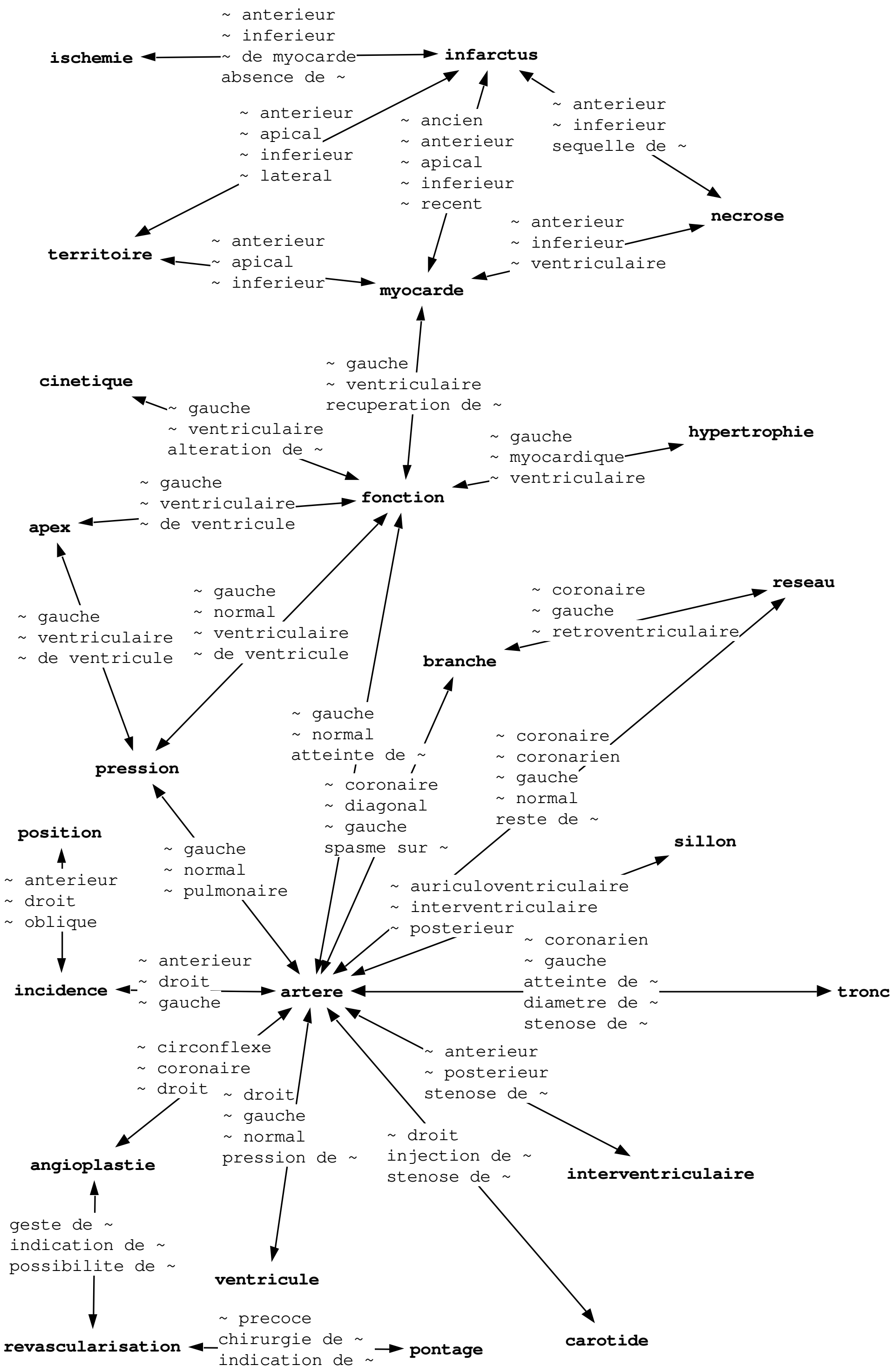

Fig. 2 - *1ère composante connexe (fin), les* **/sites/**

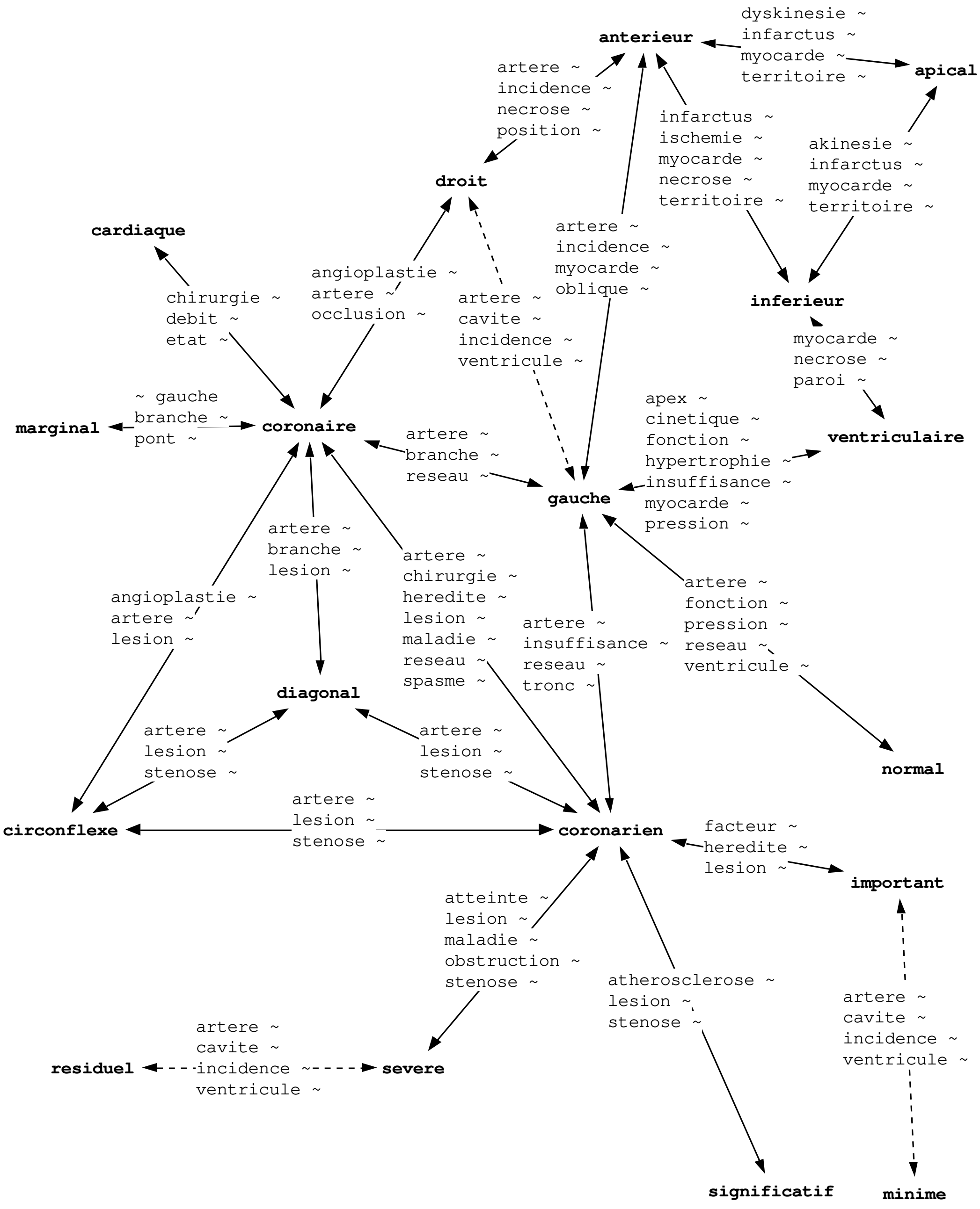

FIG. 3 - *2ème composante connexe, les adjectifs*

[Dachelet, 1994, p. 110-111][Bouaud *et al.*, 1994]. Il est donc illusoire de chercher à construire une ontologie générale. [Bouaud *et al.*, 1994] conclut d'ailleurs à l'impossibilité de concevoir une ontologie sans point de vue ou une ontologie qui ne soit pas "située". Dans cette optique, notre méthode endogène s'avère ainsi non seulement adaptée mais aussi nécessaire.

4.2 Limites d'une approche purement syntaxique

L'analyse des résultats montre que des proximités contextuelles permettent de rapprocher certains mots, mais que la syntaxe seule ne permet pas d'identifier les liens sémantiques entre ces unités lexicales. Elle suggère des liens de hiérarchisation (`générique` *vs.* `spécifique`) ou des relations conceptuelles (ex. `a pour partie`) mais des connaissances extérieures sont nécessaires pour valider la nature de ces liens. Par exemple, la similitude des syntagmes *pontage coronarien* et *pontage saphene* n'indique pas que la mention du vaisseau joue dans un cas le rôle de localisateur (pontage *sur* la coronaire) et dans l'autre celui d'un instrument (pontage *à l'aide de* la veine saphène).

De la même manière, la syntaxe ne permet pas directement de délimiter des classes de mots reflétant une notion. Les proximités contextuelles rapprochent effectivement des mots comme les noms d'`affections` (*sténose, lésion, calcification, obstruction,...*) qui se localisent sur des artères (*coronarien, circonfexe, aortique*). Mais aucune classification ne ressort directement des graphes des figures précédentes. On ne peut pas retenir les classes formées par les mots d'une composante connexe, qui seraient à l'évidence des classes trop larges. On peut repérer des zones plus denses en liens, comme *stenose, calcification, lesion, obstruction, atteinte*, mais les limites de ces zones restent floues.

Plus fondamentalement, on ne peut saisir les "existants", les référents ou notions d'un domaine, à travers les seules formes simples et les arbres élémentaires qui les relient. La simplification en arbres élémentaires a notamment le défaut de scinder des unités polylexicales (*infarctus du myocarde*) en unités lexicales élémentaires. Les termes complexes du domaine, qui renvoient pourtant à des entités, ne peuvent donc pas être situés dans le réseau construit [11].

Le fonctionnement syntaxique des unités permet donc non pas de construire une ontologie du domaine mais de dégrossir de travail de définition de cette ontologie. En ce sens, les graphes obtenus sont un outil d'exploration contextuelle. Ils ont l'avantage de fournir une vue globale des liens entre les unités alors que les concordances des lexicographes ou les indices de similarités des statisticiens

11. [Grefenstette, 1994] souligne qu'il rencontre un problème similaire. La difficulté est réelle. La délimitation des unités polylexicales échappe en effet largement à la linguistique qui ne peut donner que des indices, plus souvent négatifs que positifs, de la bonne ou mauvaise formation d'unités polylexicales potentielles [Habert and Jacquemin, 1993]. Le plus souvent, par ailleurs, on ne dispose pas de termes certifiés du domaine qui pourraient être considérés comme non décomposables dans le traitement.

[Hindle, 1990] par exemple sont toujours centrés sur une forme pôle.

4.3 Nécessité du travail interprétatif

Il est donc clair qu'un travail de classification et de structuration reste à faire pour isoler des notions et les organiser entre elles. Nous défendons l'idée que ceci relève d'un travail d'interprétation dont on ne peut faire l'économie.

Ce point de vue va à l'encontre de celui de N. Sager. Celle-ci défend l'idée que les classes de mots et les formules syntaxiques d'une grammaire d'un sous-langage sont en correspondance étroite avec les classes d'objets du monde réel et à leurs relations. Il semble cependant que le rôle de l'interprétation est très largement sous-estimé dans les travaux de l'approche harrisienne. Tout d'abord dans la construction des contextes élémentaires chez Harris [Dachelet, 1994, ch. 3]. Ces contextes résultent de simplification d'arbres comme dans SYCLADE, mais aussi de transformations. Or transformer *la sténose sévère* en *la sténose est sévère* suppose des connaissances exogènes sur *sévère* qui fonctionne comme un adjectif de degré contrairement à *coronarien* par exemple. Ensuite, on peut se demander sur quels critères autres qu'une intuition interprétative, on rapproche des critères contextuels en faisceaux de propriétés qui permettent ensuite de former des classes de mots sur la base d'une équivalence syntaxique. Enfin, c'est l'interprétation qui permet de nommer les classes de mots obtenues. Dans les travaux de l'équipe de N. Sager, il y a en fait à la fois émergence de classes en fonction du corpus et réutilisation de savoir externe [London, 1987].

Notre point de vue s'oppose également à l'idée – à notre avis illusoire – que des classes de mots sémantiquement pertinentes puissent être obtenues statistiquement et donc induites du corpus à partir des seules mesures de similarités entre mots [Church and Hanks, 1990, Hindle, 1990]. En réalité, les listes de mots obtenues ne constituent pas de véritables classes de mots cohérentes et homogènes [12]. De surcroît la cohérence sémantique, ou le sens, des classes de mots résultantes [13] n'est pas immanente, elle est le fruit d'une interprétation [14] qui est d'autant plus difficile que les mesures de similarités constituent des résumés statistiques bruts dont les critères de construction sont effacés.

Par rapport aux approches harrissiennes et statistiques, notre approche ne débouche pas davantage sur des classes sémantiques satisfaisantes. Comme pour les classes de similarité, si les composantes connexes "parlent", c'est le résultat d'un processus interprétatif inconscient et immédiat. Quelle que soit la méthode employée, la volonté d'arriver vers des classifications plus automatiques ne doit

12. Cf. [Resnik, 1993]. Seul [Grefenstette, 1994] contrôle par des calculs de similarité croisée la cohérence des classes obtenues.

13. Voici, en exemple, la "classe" des mots similaires à *treaty* obtenue par [Hindle, 1990] : *agreement, plan, constitution, contract, proposal, accord, amendment, rule, law, legislation.*

14. On trouve une confirmation de ce point dans la prudence avec laquelle [Grefenstette, 1994] exploite ses propres résultats.

pas donner l'illusion d'un sens "objectif", "machinalement" déterminé. Entre le flou notionel inhérent aux langues naturelles [Kayser, 1992], y compris aux langues de spécialités [Dachelet, 1994, p.113], et la stabilité conceptuelle qui est requise dans les ontologies construites, il y a un seuil qui ne peut être fanchi automatiquement.

Notre approche donne une bonne description des fonctionnements discursifs des unités lexicales. Mais reconnaissant le rôle central du processus interprétatif dans la constitution de catégories sémantiques ou conceptuelles, nous proposons pour le moment une structuration des résultats qui laisse explicites les critères de rapprochement entre unités et qui fournit une vue globale des données lexicales, à travers un réseau à interpréter. Elle constitue une méthode d'amorçage pour l'élaboration d'ontologies, modeste mais utile[15].

Références

[Abeillé, 1993] Anne Abeillé. *Les nouvelles syntaxes : grammaires d'unification et analyse du français.* Armand Colin, 1993.

[Bouaud *et al.*, 1994] Jacques Bouaud, Bruno Bachimont, Jean Charlet, et Pierre Zweigenbaum. Acquisition and structuring of an ontology within conceptual graphs. In *Proceedings of ICCS'94 whokshop on*Knowledge Acquisition using Conceptual Graph Theory. University of Mariland, 20-26 August 1994.

[Church and Hanks, 1990] Kenneth W. Church et Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, march 1990.

[Dachelet, 1994] Roland Dachelet. *Sur la notion de sous-langage.* Thèse de doctorat en sciences du langage, Université Paris VIII. Directeur : N. Ruwet, Décembre 1994.

[Daille, 1994] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse d'Informatique. Université de Paris VII, février 1994. Directeur: Laurence Danlos.

[Grefenstette, 1994] Gregory Grefenstette. *Exploration in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, 1994.

[Grishman and Sterling, 1992] Ralph Grishman et John Sterling. Generalizing automatyically generated selectional patterns. In *Proceedings of Coling'94*, volume 3, pages 742–747, Kyoto, 1992.

15. [Bouaud et al. 94] souligne le fait que l'on manque de méthodes pour trouver les objets et relations à y faire figurer, c'est-à-dire d'une méthodologie de l'amorçage.

[Habert and Jacquemin, 1993] Benoît Habert et Christian Jacquemin. Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *TAL*, 34(2):5–42, 1993.

[Habert *et al.*, 1996] Benoît Habert, Philippe Barbaud, Fernande Dupuis, et Christian Jacquemin. Simplifier des arbres d'analyse pour dégager les comportements syntaxico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, 1996.

[Hindle, 1990] Donnald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the Association for Computational Linguistics*, pages 268–275, 1990.

[Jacquemin, 1994] Christian Jacquemin. Recycling terms into a partial parser. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 113–118, Stuttgart, 1994.

[Kayser, 1992] Daniel Kayser. *Introduction aux sciences cognitives*, chapitre 6. Profondeur variable et sciences cognitives, pages 195–218. Coll. Folio. Gallimard, 1992.

[London, 1987] Joyce London. The healthcare lexicon. In Naomi Sager, Carol Friedman, et Margared S. Lyman, editeurs, *Medical Language Processing : Computer Management of Narrative Data*, chapitre 6, pages 137–144. Addison-Wesley, 1987.

[Resnik, 1993] Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993. (Institute for Research in Cognitive Science report IRCS-93-42).

[Ruge, 1991] Gerda Ruge. Experiments on linguistically based term associations. In *Proceedings RIAO'91*, pages 528–545, 1991.

[Sager *et al.*, 1987] Naomi Sager, Carol Friedman, et Margared S. Lyman (editors). *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley, 1987.

[Smadja, 1993] Franck Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, march 1993. Special Issue on Using Large Corpora: I.

[Zweigenbaum, 1994] Pierre Zweigenbaum. MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45:117–120, 1994.