Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles

Jacques Bouaud[†] — **Benoît Habert**[‡] — **Adeline Nazarenko**[*] — **Pierre Zweigenbaum**[†]

[†]*DIAM — SIM/AP-HP & Département de Biomathématiques, Université Paris 6*
91, boulevard de l'Hôpital, 75634 Paris cedex 13
[‡]*Équipe Linguistique et Informatique — ENS de Fontenay St Cloud*
31, avenue Lombart, 92260 Fontenay-aux-Roses
[*]*Laboratoire d'Informatique de Paris-Nord — Université de Paris-Nord*
Avenue J.-B. Clément, 93430 Villetaneuse

RÉSUMÉ.*Partant de l'expérience accumulée sur un corpus dans un domaine de spécialité (maladies coronariennes), cet article cherche à confronter deux modélisations, l'une linguistique opérée à l'aide de* ZELLIG, *un outil de regroupement des formes sur la base des dépendances syntaxiques observées, l'autre conceptuelle : la construction d'une ontologie du domaine effectuée dans le cadre du projet* MENELAS. *Il s'agit de tester l'hypothèse selon laquelle les regroupements « linguistiques » ont une pertinence ontologique et permettent d'amorcer la construction d'ontologies. La nomenclature médicale générale* SNOMED *sert également de pierre de touche. La confrontation entre ces deux approches amène à nuancer les hypothèses initiales et aide à caractériser la spécificité de chaque modélisation.*

MOTS-CLÉS : *Acquisition de connaissances à partir de textes ; Ontologie, terminologie, ingéniérie ; Validation des connaissances et systèmes à base de connaissances.*

1. Introduction

Harris [HAR 89] émet l'hypothèse qu'il est possible de mettre en évidence, à partir d'une analyse distributionnelle de contextes rendus élémentaires, les classes de concepts et les relations d'un sous-langage lié à un domaine d'activité. Cette hypothèse a été testée par Harris et son équipe à partir d'analyses syntaxiques faites à la main. Elle a aussi servi de base à des traitements automatiques de discours spécialisé [SAG 87].

ZELLIG [HAB 96a, HAB 96b] est un outil d'analyse de corpus qui s'inscrit dans la lignée des travaux de Zellig Harris et regroupe les lemmes [1] de groupes nominaux sur la base de contextes syntaxiques normalisés. Notre objectif est d'essayer d'évaluer l'apport de cet outil dans le travail de modélisation conceptuelle : des regroupements opérés sur la base de fonctionnements linguistiques communs en corpus permettent-ils l'identification et la structuration de catégories conceptuelles ?

Dans cet article, nous confrontons les catégorisations issues de ZELLIG sur des critères linguistiques à deux conceptualisations existantes dans un même domaine médical, celui des maladies coronariennes. L'expérience utilise le corpus sur ce thème rassemblé dans le cadre du projet MENELAS [ZWE 94] de compréhension de comptes rendus d'hospitalisation. Ce corpus se compose d'un extrait de manuel mais surtout de comptes rendus d'hospitalisation, ainsi que de lettres de médecins hospitaliers aux médecins traitants. Il contient 84 839 mots (occurrences) pour 6191 formes différentes.

La première catégorisation médicale que nous utilisons est la nomenclature relativement générale « SNOMED internationale » [CÔT 93] couvrant toute la médecine. La SNOMED est une hiérarchie de profondeur maximale 7. La version anglaise contient plus de 144 000 termes. Une partie de la SNOMED existe en version française, le répertoire d'anatomopathologie [CÔT 96] [2], et contient 12555 termes associés à 9097 codes (« concepts ») différents, dont 76 % sont des feuilles. La seconde catégorisation médicale est une ontologie « à grain fin » développée dans le cadre de MENELAS sur le domaine d'étude. Celle-ci est constituée de primitives de représentation structurées en arbre par la relation SORTE-DE [BOU 95, BAC 96]. Elle contient 1800 types conceptuels, avec une profondeur maximale de 20, et 71 % de feuilles [3].

La section 2 reflète le point de vue de ceux d'entre nous qui ont travaillé à la mise au point de ZELLIG. Nous y présentons brièvement les méthodes de catégorisation à partir de corpus. Nous y montrons les regroupements produits par ZELLIG. Nous y examinons les propositions de catégorisation sémantique qu'on peut faire sur cette base. Dans la section 3, rédigée par ceux d'entre nous qui ont participé à l'élaboration de l'ontologie de MENELAS, ces regroupements sont confrontés aux catégorisations conceptuelles issues de SNOMED et de MENELAS. Enfin, la section 4 et la conclusion rassemblent les deux points de vue et dégagent des pistes d'aide à la modélisation des connaissances à partir de ce type de catégorisation linguistique issue d'un travail sur corpus.

2. Catégorisation à partir de regroupements linguistiques

De nombreuses recherches, depuis une dizaine d'années, ont été consacrées à l'acquisition automatique de classes sémantiques à partir de corpus (cf. par exemple [GRE 94b]). L'approche dominante peut être divisée en trois étapes, selon [GRE 94a] : extraction des cooccurrents d'un mot, association à chaque mot de l'ensemble de ses cooccurrents et mise en évidence de la proximité/distance des mots deux à deux en fonction des cooccurrents qu'ils par-

1. Le lemme est une forme canonique qui regroupe les réalisations effectives d'un mot. On ramène ainsi les flexions d'un verbe à son infinitif, le pluriel d'un nom à son singulier, etc.

2. Nous remercions le Docteur Roger A. Côté de l'avoir gracieusement mis à notre disposition.

3. Nous remercions nos collègues du projet MENELAS, particulièrement Bruno Bachimont responsable du développement de l'ontologie.

tagent ou pas, découpage en classes en fonction des plus ou moins grandes proximités entre mots. À chaque étape, les techniques mises en œuvre varient. On peut choisir une fenêtre « graphique » (k mots de part et d'autre de chaque mot examiné) et utiliser des techniques statistiques pour regrouper les mots. On peut également utiliser un contexte syntaxique et d'autres modes de rassemblement.

2.1. Regroupements calculés à partir de dépendances syntaxiques

2.1.1. Extraction de dépendances syntaxiques

Pour rapprocher les entrées lexicales sur la base des contextes dans lesquelles elles apparaissent et dégager certaines des relations sémantiques qu'elles entretiennent, le logiciel ZELLIG [HAB 96a] s'appuie sur les arbres d'analyse produits par des extracteurs de groupes nominaux, ici Lexter [BOU 94] et AlethIPGN[4], pour en extraire les arbres élémentaires sous-jacents. Cette normalisation permet de mettre au jour les régularités que masquent les groupes nominaux complexes extraits. Pour la séquence *stenose severe de le tronc commun gauche*[5], il fournit quatre arbres élémentaires à partir de l'arbre d'analyse complet (figure 1).

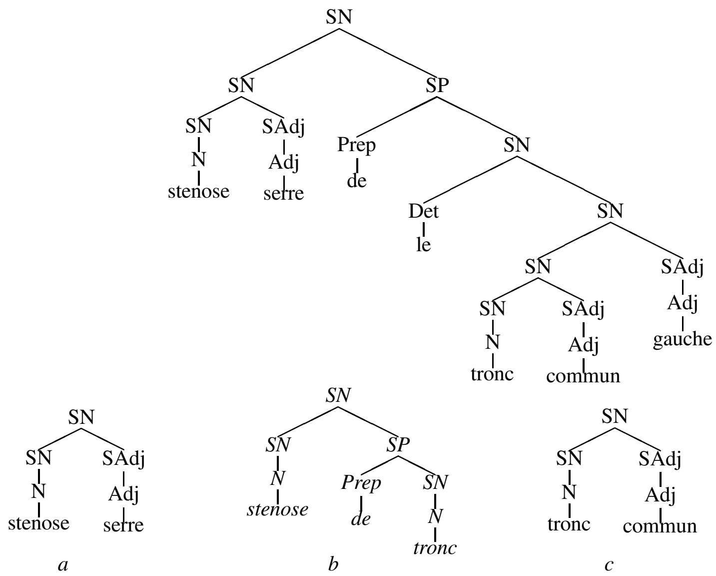

Figure 1. *Arbre d'analyse et dépendances élémentaires obtenues.*

Sont considérés comme élémentaires les arbres mettant en évidence une dépendance bi-

4. Développé par GSI-ERLI dans le cadre du projet Eureka GRAAL. Cf. [HAB 97] pour une comparaison de ces deux extracteurs.

5. Dans cette figure, comme dans tous les exemples donnés, les mots sont ramenés à leur lemme et désaccentués. Les mots contractés sont décomposés : *du* devient *de le*. C'est le résultat du programme d'étiquetage utilisé par les extracteurs dont ZELLIG retravaille les sorties.

naire entre deux mots pleins, nom ou adjectif, dans des schémas comme, par exemple, `N Prep N` ou `N Adj`. Ces dépendances associent à un élément gouverneur (nommé « tête ») soit un argument (*administration de médicament* : on administre un médicament) soit un circonstant (*administration de routine* ou *administration orale* : on l'administre comme d'habitude dans ce cas ou par voie orale). Elles ne se réalisent pas forcément par des séquences contiguës dans le corpus [6]. Mais elles correspondent à des relations de dépendance vérifiées dans les arbres d'analyse [7], si l'on passe par une représentation logique de ces arbres et des dépendances correspondantes [GAU 97].

On obtient les classes distributionnelles des mots possibles dans une position donnée d'une dépendance élémentaire, par exemple celle des mots qui commutent avec tronc dans l'arbre *b* (figure 1) : allure, artere (10 occurrences), branche (3 occ.), carotide, debut, diagonale (3 occ.), droite (4 occ.)... On peut ensuite rapprocher les formes qui partagent des contextes élémentaires (dans l'exemple : artere, carotide, branche, segment, etc. entrent dans un syntagme `N de N` où le premier `N` est stenose).

2.1.2. Calcul des regroupements par partage de contextes

ZELLIG, à partir des contextes élémentaires, construit un graphe dont les lemmes sont les nœuds et les contextes partagés par deux lemmes les arêtes. Pour limiter les proximités artificielles (liées par exemple à l'emploi de mots « vides » comme certains adjectifs très courants : certain, autre ...), un nombre minimal de contextes partagés est requis pour placer une arête entre deux lemmes : dans cette expérience et de manière arbitraire, soit 5, soit 10. Ce graphe fournit deux formes de regroupement de lemmes à travers ses composantes connexes (CC) et ses cliques (KC) [8]. Les figures [9] 2, 3 et 4 montrent respectivement la 23^{e} KC (KCA10-23 [10]) et la première CC obtenues pour AlethIPGN au seuil 10 (CCA10-1), ainsi que la deuxième CC obtenue pour Lexter au seuil 5 (CCL5-2). Au total, on obtient 30 CC (13 au seuil 5 et 17 au seuil 10) et 92 KC, toutes au seuil 10.

2.2. Examen et interprétation des regroupements linguistiques

Construire des catégories sémantiques pour une langue de spécialité est un travail laborieux, qui suppose une bonne connaissance du corpus, tout à la fois précise et globale. Nous montrons comment les graphes produits par Zellig peuvent aider des non-spécialistes dans ce travail. Les graphes sont une mise en forme du matériau textuel qui donne une vue d'ensemble

6. Par exemple, *stenose de tronc* y apparaît toujours dans des séquences plus complexes [GAU 97], et les arbres *b* et *d* de la figure 1 ne figurent pas directement dans l'arbre complet.

7. Au contraire des simples proximités textuelles, pour reprendre les critiques de [SUE 82] : */./ c'est comme si on décrivait pour chaque pièce d'un moteur, la pièce située à une proximité de 5cm, puis celle située à 10 cm, etc. Ces mesures n'auraient qu'un intérêt relatif : une pièce peut être en connexion directe avec une pièce éloignée, et fonctionner de manière totalement indépendante d'une pièce proche. Il en va de même dans les discours où les relations de dépendance syntaxique (« profondes ») entre les formes ne se confondent pas avec leurs relations de proximité (« superficielles »).*

8. Une composante connexe est une partie de graphe telle qu'il y ait un chemin entre deux nœuds quelconques. Une clique est un graphe où chaque nœud est relié à tous les autres par une arête.

9. L'outil de visualisation est GraphX (ftp://mycroft.ens-fcl.fr/pub/graphx/), développé par Serge Heiden (ELI).

10. Les CC et les KC sont suivies de l'initiale du logiciel qui les a construites (AlethIPGN/Lexter), du seuil (5 ou 10) et d'un numéro d'ordre.

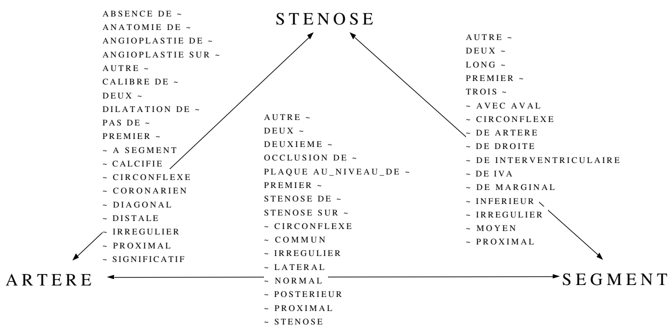

Figure 2. *La clique 23 obtenue pour AlethIPGN au seuil 10 (*KCA*10-23).*

du corpus et met en évidence le fonctionnement de certains mots dans la langue spécialisée considérée. A partir de cette cartographie, on peut, par l'examen des contextes que partagent les formes faire émerger des ébauches de catégories sémantiques. Nous montrons sur quels indices contextuels ce travail de catégorisation manuel peut s'appuyer.

2.2.1. Première cartographie du domaine

Sur les données produites par ZELLIG, quels que soient l'extracteur et le seuil utilisés, on observe une répartition similaire des résultats : une première CC pléthorique qui semble regrouper plusieurs catégories sémantiques en intersection, suivie d'une série de CC beaucoup plus petites, mais également plus homogènes.

Les premières CC mettent en évidence des lemmes « attracteurs », qui partagent des contextes avec de nombreux autres lemmes. C'est le cas pour {lesion stenose artere iva} dans CCA10-1 (figure 3). Autour de ces « attracteurs », on remarque des zones denses, enchevêtrées, où les cliques foisonnent : le groupe {lesion stenose atteinte restenose maladie pontage angioplastie artere reseau droite} autour de stenose et lesion ou le groupe {artere reseau tronc iva branche segment interventriculaire diagonale} autour de artere et iva. Enfin, des groupes de nœuds dont le rattachement au reste de la CC est ténu. L'arête angioplastie ⟷ coronarographie est le seul lien qui rattache au reste de la CCA10-1 l'étoile autour de bilan {controle epreuve exploration plan examen}, qui est homogène sémantiquement : il s'agit des examens, qui sont parfois des interventions, comme le manifeste exploration ou le lien de coronarographie avec angioplastie.

L'interprétation de ces premières CC déclenche deux types de réactions. La volonté de faire « éclater » certains regroupements qui paraissent hétérogènes, en fonction de connaissances générales sur la langue ou sur le monde : artere, stenose, angioplastie semblent relever de trois catégories distinctes : les organes, les affections au sens large (cf. le suffixe *-ose*), les interventions médicales (cf. le suffixe *-plastie* et la base *angio-* identifiant le site corporel concerné). Seconde réaction : la volonté de voir si certains regroupements, *a priori* surprenants, ne manifestent pas toutefois des comportements sémantiques propres au domaine. C'est le cas d'epreuve qui ici rentre avant tout dans le syntagme *epreuve de effort* : une série

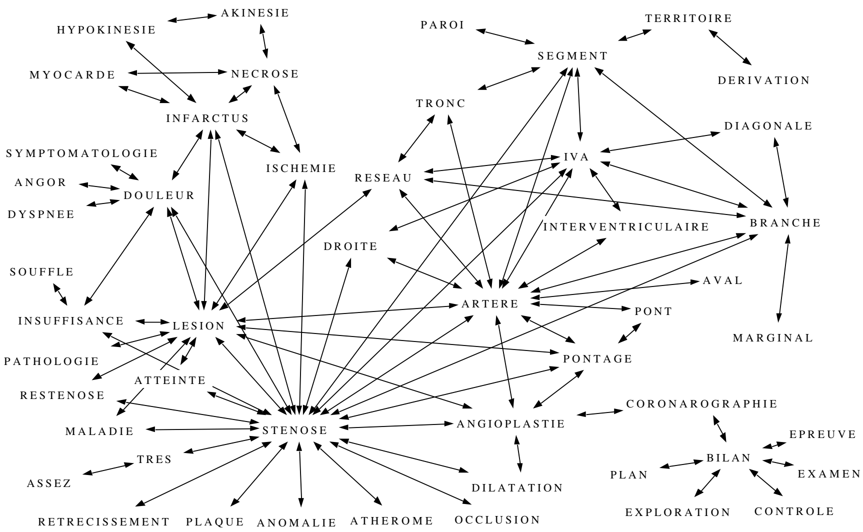

Figure 3. *La composante connexe 1 obtenue pour AlethIPGN au seuil 10 (*CCA*10-1).*

normée d'efforts demandés à un patient cardiaque à fins d'examen.

Les autres CC sont souvent réduites : entre 3 et 5 nœuds. Certaines dégagent des fragments de catégories sémantiques comme :

– la gravité d'un dysfonctionnement : {minime modere net significatif important} (CCA10-5). Une partie de cette catégorie se retrouve dans la CCL5-2 au seuil 5, avec des liens avec d'autres formes : {severe diffus} (voir plus loin, figure 4).

– la localisation au sein du myocarde ou par rapport aux affections qui le touchent : {limite anterieur posterieur lateral inferieur} où limite joue le rôle d'intrus (CCA10-7). Le seuil 5 rattache ce groupe d'adjectifs aux autres adjectifs essentiels du corpus (CCL5-2). Mais il y constitue une zone dense et homogène, d'ailleurs enrichie : {antero-apical postero-inferieur antero-lateral}.

D'autres manifestent des comportements syntaxiques particuliers. C'est le cas des adjectifs ou déterminants considérés comme des adjectifs par l'étiqueteur employé et qui peuvent se placer avant le nom : {autre premier seul deux double trois meme dernier prochain}. Il est intéressant de voir par exemple, au delà de ce qui l'est habituellement (jour, mois, semaine), que douleur est quantifié, probablement en tant que symptôme dont la fréquence et la datation comptent.

Une configuration particulière est celle des CC à trois nœuds où un nœud joue le rôle d'intermédiaire entre les deux autres. On y trouve des liens d'hyponymie, le rapprochement de synonymes et d'antonymes, ou encore l'articulation entre deux fonctionnements d'un même mot. Ces structures renseignent aussi sur des comportements spécifiques de mots courants. C'est le cas, par exemple, mauvais $\longleftrightarrow$ bon $\longleftrightarrow$ beau (CCL5-37). Dans ce type de texte, beau fonctionne comme un diagnostic positif sur les possibilités d'évolution d'un site corporel. Et hors connaissance du domaine, on aurait du mal à prédire les noms qu'il peut effectivement modifier :{calibre branche lit}. Ce qui semble d'ailleurs qualifié ici, ce n'est pas la partie du corps concernée, mais sa capacité à remplir sa fonction typique : une belle branche d'artère est une branche qui permet au sang de circuler au mieux.

2.2.2. Ébauche de catégorisations sémantiques

Une telle cartographie générale du domaine et les regroupements de formes qu'elle fait apparaître donnent des ébauches de classes sémantiques. Il reste cependant à cerner précisément les contours de ces classes et à les nommer, *i.e.* à en identifier la cohérence sémantique. Pour cette nouvelle étape, qui relève largement de l'interprétation, on peut de nouveau s'aider des graphes des contextes partagés, en prenant cette fois en compte les étiquettes des arêtes.

Notre expérience de non-spécialistes du domaine nous a montré que l'on peut exploiter les graphes issus de ZELLIG comme une carte différentielle du corpus [RAS 94]. En s'appuyant sur ces graphes, sur notre connaissance de la langue générale et sur une relative familiarité avec le corpus, on peut caractériser différentiellement les propriétés attachées à un ensemble de formes pour évaluer leur proximité sémantique.

Prenons ici l'exemple de stenose qui constitue un nœud marquant de CCA10-1 (figure 3). Stenose, situé au centre d'une étoile très fournie, a 22 voisins : atherome, maladie, tres, retrecissement, plaque, anomalie, dilatation, occlusion, angioplastie, pontage, branche, IVA, artere, segment, droite, infarctus, ischemie, douleur, insuffisance, lesion, atteinte, restenose. Les nombreuses cliques dans lesquelles entre stenose suggèrent par ailleurs que {stenose lesion douleur insuffisance infarctus} constituent une famille de mots assez homogène.

L'examen des contextes d'une forme donnée et leur typage permet de décrire les attributs qui caractérisent le concept auquel elle renvoie et les relations dans lesquelles il entre. On peut par exemple regrouper manuellement les modifieurs de degré ou les adjectifs susceptibles de modifier myocarde... Dans le cas de stenose, ce travail assez grossier de typage permet d'identifier 5 propriétés. Une sténose se caractérise par son aspect – sa longueur et son diamètre, notamment – (P_1), par son degré de gravité (P_2) et par sa localisation, en particulier sur une artère ou une partie d'artère (P_3). Enfin, une sténose fait l'objet d'un acte thérapeutique (P_4) et peut être considérée comme un processus (P_5).

Si l'on considère alors les contextes que stenose partage avec chacun de ses voisins, on peut établir le classement suivant.

1. Seul lesion partage avec stenose tous ses types de contextes : la forte proximité entre ces deux lemmes est confirmée à la fois par le nombre de contextes partagés, par le nombre des occurrences de contextes concernés et par le nombre de types de contextes partagés, *i.e.* par la diversité sémantique de ces contextes.

2. De nombreux lemmes partagent avec stenose plusieurs types de contextes (3 ou 4) mais pas tous. Parmi cet ensemble certaines sous-catégories se dégagent :

   (a) 5 lemmes {insuffisance atteinte maladie plaque douleur} ont en commun de partager toutes les propriétés de stenose, sauf P_4. À l'exception de plaque, ils fonctionnent tous dans le corpus comme des hyperonymes de stenose en désignant des pathologies de manière plus générique, à partir d'un diagnostic moins précis.

(b) ischemie et infarctus se distinguent de l'ensemble : ils ne peuvent être localisés que par rapport au cœur. Ils s'écartent de ce fait de la famille de stenose pour laquelle la localisation artérielle est centrale (P_3).

(c) 4 lemmes {IVA artere droite branche} se rapprochent de stenose du fait de P_4 et P_1, mais surtout de P_3 (nombre important à la fois de contextes et d'occurrences de contextes).

(d) 2 lemmes {occlusion restenose} partagent les propriétés P_1, P_2, P_3 et P_4. Ils se distinguent de artere et droite mentionnés ci-dessus par le fait que, cette fois, P_1 et P_2 jouent un rôle important.

(e) 5 lemmes {anomalie atherome retrecissement dilatation tres} se rapprochent de stenose sur la base de ses trois attributs principaux : P_1, P_2 et P_3.

3. 3 lemmes {pontage angioplastie segment} ne partagent que deux attributs avec stenose. Dans ces cas-là, le nombre de contextes différents fait apparaître une proximité sur les graphes, mais ces contextes sont tous du même type et cette proximité reste en fait partielle [11].

Sur la base de cette analyse, on peut proposer de construire des familles sémantiques. On voit en effet se dessiner la famille des pathologies (1, 2d) et des pathologies artérielles (2a). Les groupes 2b, 2c et 3 se distinguent nettement de celle-ci. Il s'agit de trois groupes homogènes : les pathologies cardiaques, les artères ou parties d'artères et les actes thérapeutiques. Tous ces regroupements sont par ailleurs confirmés par des CC/KC. Reste le groupe 2e qui semble hétérogène et difficile à classer. On y trouve 3 pathologies {anomalie atherome retrecissement}, un acte et l'adverbe tres. C'est l'attribut du degré, possible pour dilatation mais pas pour pontage ou angioplastie, qui le distingue du groupe 3 dans lequel il semble pourtant se ranger.

Notre hypothèse de départ trouve un début de confirmation. L'étude des graphes et des contextes partagés entre formes aide à construire des classes sémantiques. Elle permet d'étudier les emplois et les proximités entre formes dans une langue de spécialité, même mal connue, en s'affranchissant des préjugés induits par la langue générale.

3. Confrontation à deux modélisations conceptuelles

Notre objectif est ici d'examiner dans quelle mesure le résultat de ZELLIG est compatible avec ce que l'on peut trouver dans une modélisation conceptuelle existante. Il ne s'agit pas de la validation d'une méthode d'acquisition de hiérarchie conceptuelle reposant sur ZELLIG, mais plutôt d'une première exploration visant à mettre au point une telle méthode. Nous décrivons ailleurs comment on pourrait utiliser ZELLIG pour adapter ou raffiner une modélisation existante pour un corpus donné [NAZ 97].

Nous considérons la hiérarchie de concepts présente dans une modélisation conceptuelle. Nous supposons de plus qu'un « lexique sémantique » fait le pont entre mots de la langue

11. Sur les 14 contextes que pontage et angioplastie partagent avec stenose, 9 ou 10 introduisent une localisation sur une artère; 10 des 16 contextes partagés entre segment et stenose introduisent une localisation.

(nos lemmes) et les catégories conceptuelles du système de référence. Ainsi, chaque lemme présent dans le lexique possède un ou plusieurs types. La question que nous nous posons est alors : est-ce que les lemmes voisins dans les graphes de ZELLIG possèdent des types proches, voire identiques selon cette hiérarchie ? Et par conséquent, est-ce que les CC et KC regroupent des lemmes de types proches, voire identiques ? Pour répondre à cette question, nous remplaçons dans les graphes de ZELLIG chaque lemme par son type, et examinons les propriétés des graphes résultants.

Nous avons utilisé dans ce but deux modélisations existantes, de nature très différente. La première, la nomenclature « SNOMED internationale » [CÔT 93], est une catégorisation très employée dans le domaine médical. La SNOMED est la terminologie médicale qui possède le meilleur compromis entre structuration (hiérarchie à 7 niveaux), couverture et complétude (144 000 termes), et finesse de grain (non seulement classifications de diagnostics et d'actes, mais également nomenclatures d'anatomie, signes et symptômes, produits chimiques, etc.). La SNOMED occupe une position intermédiaire entre un thésaurus (linguistique) et un système formel de notions (ontologie) [GAN 92]. La seconde, l'ontologie de MENELAS [BOU 95, BAC 96], a été conçue dans le cadre d'un projet multilingue d'analyse de textes, comme fondement d'une représentation conceptuelle normalisée devant servir de support à des inférences indépendantes de la langue. Bien que d'une complétude moindre que la SNOMED (1800 types conceptuels), elle est davantage structurée (principes de structuration ; arbre de profondeur maximale 20). Le domaine médical visé est celui des maladies coronaires, mais le haut de la hiérarchie conceptuelle déborde largement le domaine médical.

3.1. Confrontation à une catégorisation « à gros grain » : « SNM »

La SNOMED couvre essentiellement le vocabulaire biomédical, avec une ouverture vers des termes plus généraux (classe G: qualificatifs et termes relationnels). Par ailleurs, la SNOMED contient des termes généralement composés de plusieurs mots, alors que ZELLIG fonctionne (dans l'immédiat) sur des mots simples. Nous avons donc adapté la SNOMED en travaillant sur ses 11 classes les plus générales et en ajoutant 9 classes supplémentaires, dont la plupart sous-catégorisent la classe G de la SNOMED. Nous désignerons par « SNM » cette catégorisation. 8 des 17 classes de SNM sont directement issues de la SNOMED (A C D F M P S T : Agents et Artefacts, produits Chimiques et biologiques, Diagnostics et maladies, Fonction, Morphologie, Procédures, contexte Social, Topographie). Pour des raisons pratiques, nous avons travaillé sur les lemmes présents dans le lexique sémantique de MENELAS, qui possède une bonne couverture des mots les plus fréquents dans le corpus ; nous avons ainsi catégorisé 845 lemmes simples selon les 17 classes de SNM. Nous appellerons « lexique SNM » le lexique sémantique résultant.

3.1.1. Homogénéité sémantique globale des CC

Si l'on considère l'ensemble des 30 CC trouvées, on observe après substitution que l'étiquetage est homogène pour 18 d'entre elles, 3 contiennent 2 classes très proches, et 2 sont inhomogènes, mais du fait d'incohérences dans la constitution du lexique SNM. Au total donc, 23 CC sur 30 confirment notre attente.

Parmi les 7 restantes, 5 sont grosses à très grosses (14, 34, 50, 99, 233 lemmes). Nous les examinons plus loin. Les 2 dernières contiennent un « intrus » comme par exemple CCA10-7 :

{posterieur/GT anterieur/GT apical/ inferieur/GT lateral/GT limite/}

Ici, limite est hors du lexique SNM. L'examen du corpus montre que ce terme ne désigne pas une localisation relative (GT : topographie générique). Il s'agit sans doute d'un modifieur général (MA), qui apparaît surtout dans les locutions « à la limite de » et « dans les limites de ». Nous avons donc ici un véritable intrus, qui se trouve partager un certain nombre de contextes avec anterieur. On peut noter cependant que sur le graphe, {posterieur/GT anterieur/GT apical/ inferieur/GT lateral/GT} forment une clique, dont limite ne fait pas partie.

Revenons maintenant aux « grosses » CC, qui contiennent plusieurs catégories SNM, et ne peuvent pas être abordées de la même façon que les petites CC uniformes. Nous essayons d'exploiter leur structure graphique, en tenant compte de la connectivité des nœuds. Nous supposons que les nœuds d'une même classe sémantique peuvent être regroupés dans des sous-parties du graphe fortement connexes mais relativement disjointes du reste du graphe. Ce principe, appliqué aux 5 CC concernées, montre des regroupements locaux. Ainsi, CCL5-2 (34 lemmes) regroupe (voir figure 4) :

GT (localisations relatives) : {gauche droit anterieur inferieur lateral antero-lateral posterieur apical postero-inferieur antero-apical} {moyen proximal distale}

MA (qualificatifs) : {important significatif severe minime recent ancien}

T (topographie) : {coronaire coronarien cardiaque circonflexe aortique mitral valvulaire}

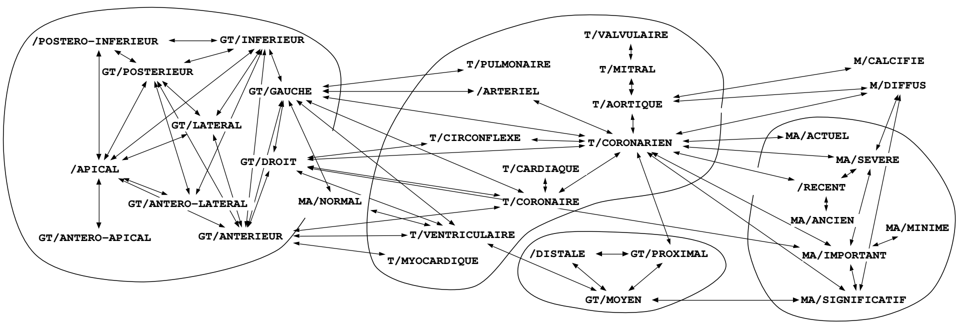

Figure 4. *Regroupements de lemmes de même catégorie SNM dans* CCL*5-2. Les lemmes présents dans le lexique SNM sont précédés d'une étiquette (ex.* T/MITRAL*), les autres n'en comportent pas (ex.* /ARTERIEL*)*

En résumé, si l'on élimine la plus grosse CC [12], les 29 CC restantes permettent de déterminer 37 groupes homogènes dont plusieurs se recoupent. Certains d'entre eux correspondent à des synonymies, comme {coronarien coronaire} (CCA10-3). Quand on considère les KC, en revanche, les résultats sont plus mitigés : 68 sur 92 sont inhomogènes.

12. La plus grosse CC est en effet trop volumineuse pour être affichée par notre outil de visualisation.

3.1.2. Exploitation des groupements pour la sous-catégorisation

On peut affiner les classes à gros grain de la SNM. Certaines donnent en effet lieu à plusieurs sous-groupes : dans CCL5-1, on a deux (voire trois) sous-groupes de procédures (P), reliées l'une à l'autre par une seule arête, et possédant des liens avec des groupes différents de lemmes de classes autres que P. La distinction en deux sous-groupes se retrouve dans CCA10-1. En faisant l'union des deux, on obtient :

P (procédure) : {bilan exploration controle examen epreuve test coronarographie plan}
{traitement therapeutique}
{angioplastie dilatation revascularisation pontage intervention hospitalisation}

où l'on reconnaît les examens, les traitements (médicamenteux), et les traitements invasifs (sauf hospitalisation). De même, les localisations relatives (GT) se divisent sur CCL5-2 en {gauche droit anterieur inferieur lateral antero-lateral posterieur apical postero-inferieur antero-apical} et {moyen proximal distale}, que l'on peut interpréter comme des localisations cardiaques et artérielles, respectivement. On trouve également une sous-classe de MA (qualificatifs) : {important significatif severe minime recent ancien}, modifiant essentiellement les problèmes coronariens.

De façon générale, chaque groupe homogène identifié constitue potentiellement une sous-classe de la catégorie SNM commune à ses lemmes. Par exemple, le groupe des VE {existence absence presence aspect pas recidive} (CCL5-1) pourrait constituer des indicateurs du mode d'existence d'une observation, et s'opposer aux autres VE (p.ex., {modification elevation alteration trouble} (CCL5-1), que l'on peut interpréter comme des modifications d'un état). Certains groupes sont inclus dans d'autres : {mitral aortique} (CCA10-8) est inclus dans {coronaire coronarien cardiaque circonflexe mitral aortique valvulaire} (CCL5-2), et pourrait en constituer une sous-classe : « éléments qui possèdent une valve » $\subset$ « éléments du système cardio-vasculaire » $\subset$ « parties du corps » (T). De même, {absence pas} (CCA10-10) est inclus dans les indicateurs de mode d'existence, et regroupe les expressions de non-existence. Là où il était possible, l'examen de la hiérarchie complète de la SNOMED est en accord avec les sous-classes trouvées [NAZ 97]. Par exemple, la SNOMED possède des sous-classes de la catégorie P (Procédures) pour les examens (P3–P5), les traitements médicamenteux (P2) et les traitements invasifs (P1).

Les groupes sont cependant en général séparés par la catégorie lexicale des lemmes : les noms de topographie {branche artere reseau iva interventriculaire diagonale tronc segment paroi pont} forment une hiérarchie différente des adjectifs de topographie qui viennent d'être présentés.

3.2. Une catégorisation fine : l'ontologie de MENELAS

3.2.1. Modélisation conceptuelle dans MENELAS

Une partie importante des connaissances de MENELAS réside dans son *ontologie* qui répertorie et organise les notions du domaine. En dehors de toute modélisation formelle *a priori*, l'élaboration d'une ontologie, noyau de tout système de représentation, passe par une étape d'acquisition pour laquelle on dispose de peu de guides [GRU 95, BAC 96]. Dans les domaines non formels, comme la médecine, l'accès aux connaissances passe essentiellement par leur expression en langue naturelle ; il n'existe pas de définitions complètes, nécessairement uniques et non contextuelles des notions mobilisées. Aussi, à l'occasion du projet MENELAS, une étape de normalisation et des principes de structuration ont-ils été proposés [BAC 96, BOU 95] afin de stabiliser le sens des notions. À l'issue de cette étape, chaque notion ne s'interprète plus par le biais de son label, *i.e.*, en langue, mais par la suite unique des propriétés qui le définissent, *i.e.*, par la place qu'il occupe dans l'ontologie.

L'acquisition de cette ontologie, qui se compose d'environ 1800 types et 300 relations, a été effectuée selon la méthodologie évoquée [BOU 95, BAC 96] et en exploitant plusieurs sources : des entretiens avec des médecins, la consultation de dictionnaires médicaux (p. ex. [GAR 92]), un essai de réutilisation de ressources terminologiques existantes [VOL 93], l'introspection et surtout l'examen d'une partie du corpus constitué sur ce thème. Cependant, l'accès au corpus n'a pas été systématique, en partie du fait du manque d'outils évolués au-delà d'un simple concordancier.

3.2.2. Principes d'examen des regroupements

Dans MENELAS, les lemmes renvoient, via le lexique sémantique, à des types ontologiques généralement distincts, ce qui ne permet pas, après substitution dans les CC/KC, d'observer directement une homogénéité d'étiquetage comme ce pouvait être le cas avec SNM. La question ici est alors de savoir dans quelle mesure les types ramenés pour une même CC/KC partagent une communauté de sens (outre celle d'exister dans le domaine). Les types étant hiérarchisés, on recherche pour chaque graphe de ZELLIG le surtype commun minimal subsumant tous les types des sommets [13], ou « type subsumant ». On cherche ensuite si ce surtype correspond ou non à une catégorie ontologique « générale ». Bien que la profondeur ne soit pas un indicateur homogène de distance sémantique entre types, nous avons arbitrairement considéré comme « généraux » les types d'une profondeur strictement inférieure à 4 par rapport à la racine. Ces types correspondent grossièrement aux *objets intentionnels* (*états* et *actions*), aux *objets physiques*, aux *notions* (*mathématiques, médicales*) et aux *points de vue*.

Certaines CC/KC sont inexploitables du fait de lacunes du lexique de MENELAS [14]. Dans la moitié des cas, les types subsumant les CC/KC sont très généraux, traduisant des regroupements ontologiquement hétérogènes alors que l'autre moitié reflète une cohérence conceptuelle plus spécifique. Nous décrivons et discutons ci-dessous certains cas de figure rencontrés.

3.2.3. Regroupements ontologiquement « spécifiques »

On considère ici que les lemmes partagent effectivement un sens commun spécifique. On retrouve ainsi la classe des localisants par rapport au cœur : les lemmes {anterieur apical inferieur lateral posterieur limite} (CCA10-7) renvoient les types conceptuels {ANTERIOR APICAL INFERIOR LATERAL POSTERIOR} [15] dont le type subsumant minimal est RELATIVE_LOCATION (*localisation relative à un système de repérage*). La cohérence conceptuelle de cette classe de lemmes est confirmée dans 11 autres KC qui ramènent également ce type pour le même

13. On notera que, pour tout ensemble de types, il est toujours possible de trouver un tel surtype : au plus la racine de l'arbre.
14. Toutefois, 98 % des 303 sommets des KC et 71 % des 507 sommets des CC sont couverts.
15. Limite, n'étant pas défini dans le lexique, est ignoré.

groupe d'adjectifs. Cependant, dans l'ontologie, ces types ne sont pas tous sous-types de RELATIVE_LOCATION de la même manière. Il est intéressant de noter que ceci est (en partie) montré par 5 autres KC qui font apparaître le sous-ensemble plus spécifique {anterieur apical inferieur posterieur} qui renvoie au type VOLUMIC_RELATIVE_LOCATION plus spécifique que RELATIVE_LOCATION. Dans les 11 KC précédentes l'« intrus » est lateral, qui renvoie à PLANAR_RELATIVE_LOCATION (non pas comme les autres à VOLUMIC_RELATIVE_LOCATION), forçant la généralisation à leur surtype commun RELATIVE_LOCATION.

Le cas des adjectifs {mitral aortique}, qui signifient respectivement « *en forme de mitre* » et « *relatif à l'aorte* » et qui a priori ont peu en commun, est singulier car ils sont regroupés par ZELLIG (CCA10-8 et CCL10-5). Pourtant, cette communauté de comportement linguistique est dans un premier temps confirmée par l'ontologie avec comme type subsumant non général ANATOMICAL_ELEMENT. On voit ici la spécificité du domaine : mitral est dans le corpus toujours en rapport avec la *valve mitrale* ; il l'est également dans l'ontologie du domaine et renvoie donc au type MITRAL_VALVE. Il existe également une *valve aortique* et c'est bien en tant que qualifieurs (directs ou indirects) de valves que ces adjectifs partagent des contextes. On aurait donc pu s'attendre à trouver VALVE comme type subsumant. Mais aortique ne désigne pas de manière univoque la *valve aortique* et peut qualifier d'autres lemmes même dans le domaine (p. ex. *crosse aortique*). Aussi, aortique conserve-t-il dans l'ontologie son sens général et renvoie au type AORTA (*l'aorte*), bien que le type AORTIC_SIGMOID_VALVE (*valve aortique*) existe. Ainsi, *artères* et *valves* sont classés distinctement dans l'ontologie, même s'ils ont pour surtype commun minimal ANATOMICAL_ELEMENT.

En ce qui concerne les KC, celles ramenant des types spécifiques sont peu nombreuses. Outre RELATIVE_LOCATION mentionné plus haut, on y trouve le type STATE_OF_PHYSICAL_OBJECT (23 KC) regroupant les lemmes comme {stenose lesion douleur ischemie infarctus necrose} mais qui reste néanmoins un type assez général. Plus spécifiquement, on trouve pour {stenose restenose lesion} (KCL10-27) le type MORPHO_ANOMALY qui subsume directement STENOSIS. Les lemmes stenose et restenose renvoient à ce dernier alors que lesion renvoie au premier. L'ontologie reflète le fait que lesion a été considéré comme hyperonyme de stenose, cependant, cette subtilité ne se retrouve pas directement à l'examen des KC.

3.2.4. Regroupements ontologiquement « généraux »

Les CC/KC ramenant des types subsumants généraux montrent que la communauté de comportement linguistique n'implique pas nécessairement une communauté conceptuelle. Cette caractéristique était attendue pour les grandes CC. On notera toutefois au passage que sur CCL5-2 on retrouve, mais à un niveau assez haut dans l'ontologie, les groupements obtenus avec SNM (figure 4). Sur cette CC les étiquetages SNM GT et T correspondent exactement à RELATIVE_LOCATION et à ANATOMICAL_ELEMENT, (GT $\cup$ MA) à VIEW_POINT. En revanche, cette propriété était moins attendue pour les KC, les petites notamment.

Ainsi, le type subsumant auquel renvoient les lemmes {patient monsieur} (CCA10-6) est SUBSTRATUM, racine de l'ontologie. Dans l'ontologie, les types PATIENT et HUMAN_BEING correspondants sont bien distincts : le premier est un *rôle* alors que le second est un *être animé*. Pourtant, dans le corpus, ces deux lemmes sont pratiquement substituables. Les rapprochements obtenus mettent en évidence des figures de style comme la métonymie [16] : une personne (le patient) est désignée indifféremment par son essence ou par son rôle, tout du moins dans de nombreux contextes.

La situation est similaire dans KCA10-23 {artere stenose segment} (voir figure 2) pour laquelle on obtient encore SUBSTRATUM. Dans l'ontologie, ARTERY est un *objet anatomique*, STENOSIS un *état* et SEGMENT un *objet géométrique*, types qui ne manifestent aucune communauté conceptuelle. Les connaissances encyclopédiques rattachées à ces notions montrent, en revanche, que ces trois notions sont intimement liées. En effet, une *sténose* est caractérisée par un rétrécissement du diamètre d'une *artère* localisée sur un *segment* de cette artère. Ainsi, en considérant une *dilatation* qui est une *intervention* visant à élargir le diamètre d'une *artère* rétréci sur un *segment*, les séquences *dilatation d'artère*, *dilatation de sténose* et *dilatation de segment* signifient en pratique la même chose en ce qu'elles conduisent à faire les mêmes inférences. Dans le contexte d'une dilatation, les trois lemmes sont ici métonymes, même si l'accent ne porte pas sur le même concept. Dans le cadre d'un glissement de sens métonymique, les trois lemmes doivent être ici interprétés avec le même sens. Le regroupement linguistique trouvé par ZELLIG traduit bien alors dans ce cas une communauté de sens. Seulement, cette communauté est simplement locale et n'a pas de valeur en dehors de ces contextes-là ; elle ne reflète pas le point de vue ontologique qui, lui, vise à s'affranchir du contexte.

4. Les regroupements en corpus : une aide à la construction d'ontologies

4.1. Un outil de débroussaillage

Les classifications linguistiques opérées dans une optique harrissienne, c'est-à-dire la construction de classes d'équivalence distributionnelle sur la base de contextes rendus élémentaires, débouchent sur des regroupements relativement généraux mais dont la confrontation avec la catégorisation SNOMED montre la pertinence ontologique. Les proximités manifestées par les CC ou les KC peuvent donc aider à la mise au point d'une première classification, grossière. On a ainsi isolé quelques embryons de classes relativement homogènes : un groupe autour de stenose, un ensemble d'adjectifs de degré, un ensemble de localisants.

La confrontation entre les deux approches de modélisation amène d'ailleurs à vouloir se baser plutôt sur les CC que sur les KC. Ceci nous amène à nuancer les conclusions de [HAB 96b]. Il semble en effet que les rapprochements opérés par les CC sont relativement pertinents du point de vue ontologique, qu'il s'agisse de l'ensemble des lemmes – pour les petites CC –, ou de sous-ensembles – pour les plus grosses. Les CC peuvent donc être exploitées pour contruire une première catégorisation. À l'inverse, les KC, plus hétérogènes, font apparaître les frontières entre classes et les articulations entre les différents sens des lemmes. Elles rapprochent des lemmes qui peuvent renvoyer à des concepts très différents en raison de phénomènes de glissements de sens, où un mot change de catégorie conceptuelle.

16. Procédé de langage par lequel on exprime un concept au moyen d'un terme désignant un autre concept uni par une relation nécessaire.

4.2. *Deux modélisations irréductibles mais complémentaires*

La confrontation avec la catégorisation de MENELAS montre aussi les limites des regroupements opérés par les CC/KC pour l'élaboration de classes fines. L'étude des contextes permet néanmoins d'affiner certaines classes et d'en décrire les éléments avec précision, comme dans le cas du groupe des localisants qui peut se décomposer en sous-groupes. On voit alors se dessiner une méthode de travail incrémentale qui repose sur un processus en spirale et des allers-retours répétés entre les rapprochements opérés par les graphes et l'observation des contextes élémentaires des formes. À partir de classes grossières, on peut identifier les attributs et propriétés du concept auquel renvoie stenose. Les premières classes construites servent à en élaborer de nouvelles et l'étude de ces dernières permet en retour d'affiner les premières. Ce versant linguistique du travail ne peut être entièrement automatisé. Il s'appuie sur la connaissance de la langue générale (utilisation d'informations sémantiques apportées par les préfixes et suffixes par exemple) et sur une part incontournable d'interprétation.

Nous ne pensons pas pour autant que ce processus en spirale puisse déboucher directement sur les primitives d'une ontologie. Les regroupements sur la base de dépendances syntaxiques ne débouchent pas forcément sur des types disjoints et structurés. Les contextes qui fournissent ces regroupements permettent seulement partiellement de dégager les propriétés caractéristiques des mots rassemblés. La confrontation des deux modélisations manifeste en fait le fossé entre les catégorisations floues et grossières dont s'accommode la langue et le raffinement et la précision nécessaires à une ontologie formelle.

5. Conclusion

Nous nous garderons d'élargir à l'approche harrissienne tout entière les limites que nous avons pu constater lors de notre examen des résultats de ZELLIG. ZELLIG s'écarte en effet de la méthodologie harrissienne sur trois points cruciaux. Tout d'abord, ses regroupements, faute de pouvoir, dans l'immédiat, recourir à des analyseurs de *phrase* fiables, se cantonnent aux groupes nominaux. Même si les nominalisations [17] tiennent une place importante dans le corpus [18], les dépendances sujet/verbe, verbe/circonstant et verbe/objet permettraient de mettre au jour d'autres patrons informationnels fondamentaux. Ensuite, la régularisation des contextes pourrait utiliser des transformations syntaxiques plus complexes : par exemple *sévérité de sténose* pourrait être ramené à *sténose sévère*. Enfin, Harris et son équipe avaient parfois recours à des experts du domaine pour l'obtention de certaines catégories « élémentaires » [DAL 90, p.75], alors qu'avec ZELLIG, le corpus et les regroupements qu'il permet constituent la seule source d'information.

Nous avons exploré une petite partie seulement des informations provenant des dépendances « normalisées » produites par ZELLIG (le nombre d'occurrences des dépendances n'a pas été exploité, par exemple). D'autres modes de regroupement sont en outre envisageables, en faisant appel en particulier aux techniques statistiques de classification automatique [LEB 94]. La réduction des proximités artificielles passe, elle, par la mise en évidence

17. Par exemple, *dilatation de l'artère circonflexe* pour *on dilate l'artère circonflexe*.
18. Les groupes nominaux extraits par Lexter at AlethIPGN couvrent respectivement 27 et 38 % du corpus.

puis l'élimination des contextes introduisant du bruit (ce qui varie d'un corpus à l'autre). L'utilisation de seuils dans la présente expérience n'en constitue qu'une approximation grossière.

Les outils d'exploration et de fouille de corpus se multiplient. Le bilan de leur apport progresse plus lentement. Il est trop coûteux en effet de mettre sur pied un protococole d'évaluation rigoureux, qui pourrait être basé dans le cas présent sur la construction en parallèle d'une ontologie avec et sans l'aide de ce type d'outils. Si l'expérience présente ne constitue pas un tel protocole [19], elle confronte néanmoins les deux points de vue distincts entre lesquels les auteurs de l'article se répartissent : la catégorisation à partir de regroupements linguistiques (section 2) et la mise au point de modélisations conceptuelles (section 3). Cependant elle porte avant tout sur la pertinence pour une modélisation conceptuelle de regroupements issus de dépendances syntaxiques élémentaires. Il reste à examiner en retour si ces regroupements et les informations qui leur sont associées pourraient amener à remanier de manière significative une ontologie comme celle de MENELAS.

Références

[BAC 96] BACHIMONT B., *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser ; Critique du formalisme en intelligence artificielle*, Thèse de doctorat d'épistémologie, École Polytechnique, 1996.

[BOU 95] BOUAUD J., BACHIMONT B., CHARLET J. et ZWEIGENBAUM P., « Methodological principles for structuring an "ontology" », *Proceedings of the IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing."*, Montréal, 1995.

[BOU 94] BOURIGAULT D., *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*, Paris, Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, 1994.

[CÔT 93] CÔTÉ R. A., ROTHWELL D. J., PALOTAY J. L., BECKETT R. S. et BROCHU L. (édité par), *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*, Northfield, College of American Pathologists, 1993.

[CÔT 96] CÔTÉ R. A., *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*, Université de Sherbrooke, Sherbrooke, Québec, 1996.

[DAL 90] DALADIER A., « Aspects constructifs des grammaires de Harris », *Langages*, n° 99, p. 57–84, 1990, A. Daladier (resp.).

[GAN 92] GANGEMI A., GALANTI M., GALEAZZI E. et ROSSI MORI A., « Beyond UMLS: Computational semantics for medical records », *Proc MEDINFO 92* éd. par LUN K. C., DEGOULET P., PIEMME T. et RIENHOFF O., Geneva, p. 703–708, 1992.

[GAR 92] GARNIER M. et DELAMARE V., *Dictionnaire des Termes de Médecine*, Paris, Maloine, 1992.

19. En particulier parce que nous n'avons pas découvert le corpus à l'occasion de cette confrontation : notre familiarité avec le domaine et sa modélisation biaise notre bilan.

[GAU 97] GAUSSIER E. et HABERT B., « Langue spécialisée : des séquences observées aux mots possibles », *Mots possibles et mots existants* éd. par CORBIN D., FRADIN B., HABERT B., KERLEROUX F. et PLÉNAT M., Lille, 1997.

[GRE 94a] GREFENSTETTE G., « Corpus-derived first, second and third-order word affinities », *EURALEX*, Amsterdam, 1994.

[GRE 94b] GREFENSTETTE G., *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers, 1994, *Kluwer international series in engineering and computer science*.

[GRU 95] GRUBER T. R., « Toward principles for the design of ontologies used for knowledge sharing », *International Journal of Human-Computer Studies*, vol. 43, n ° 5/6, p. 907–928, 1995, Special issue on ontology.

[HAB 97] HABERT B., HERVIOU-PICARD M.-L., BOURIGAULT D., QUATRAIN R. et ROUMENS M., « Un outil et une méthode pour comparer deux extracteurs de groupes nominaux », *1ères Journées Scientifiques et Techniques FRANCIL*, Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, Avignon, 1997.

[HAB 96a] HABERT B., NAULLEAU E. et NAZARENKO A., « Symbolic word clustering for medium-size corpora », *16th International Conference on Computational Linguistics*, Copenhagen, Denmark, p. 490–495, 1996.

[HAB 96b] HABERT B. et NAZARENKO A., « La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience », *Journées sur l'acquisition des connaissances*, AFIA, Sète, p. 137–142, 1996.

[HAR 89] HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK JR P., DALADIER A., HARRIS T. et HARRIS S., *The Form of Information in Science, Analysis of Immunology Sublanguage*, Dordrecht, The Netherlands, Kluwer Academic Publisher, 1989, *Boston Studies in the Philosophy of Science*, volume 104.

[LEB 94] LEBART L. et SALEM A., *Statistique textuelle*, Paris, Dunod, 1994.

[NAZ 97] NAZARENKO A., ZWEIGENBAUM P., BOUAUD J. et HABERT B., « Corpus-based identification and refinement of semantic classes », *AMIA Annual Fall Symposium 97* éd. par MASYS D. R., AMIA, 1997, Submitted.

[RAS 94] RASTIER F., CAVAZZA M. et ABEILLÉ A., *Sémantique pour l'analyse : de la linguistique à l'informatique*, Paris, Masson, 1994, *Coll. Recherches cognitives*.

[SAG 87] SAGER N., FRIEDMAN C. et LYMAN M. S. (édité par), *Medical Information Processing — Computer Management of Narrative Data*, Reading, Mass, Addison Wesley, 1987.

[SUE 82] SUEUR J.-P., « Pour une grammaire du discours : élaboration d'une méthode; exemples d'application », *MOTS*, n ° 5, p. 145–185, 1982.

[VOL 93] VOLOT F., ZWEIGENBAUM P., BACHIMONT B., BEN SAÏD M., BOUAUD J., FIESCHI M. et BOISVIEUX J.-F., « Structuration and acquisition of medical knowledge: Using UMLS in the Conceptual Graph formalism », *Proceedings of the 17th Annual SCAMC*, Washington, p. 710–714, 1993.

[ZWE 94] ZWEIGENBAUM P. et CONSORTIUM MENELAS, « MENELAS: an access system for medical records using natural language », *Computer Methods and Programs in Biomedicine*, vol. 45, p. 117–120, 1994.