Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies ⋆

Pierre Zweigenbaum †, **Jacques Bouaud** †, **Benoît Habert** ‡ et **Adeline Nazarenko** ⋆
† DIAM — SIM/AP-HP & Dépt. Biomathématiques U. Paris 6
‡ Équipe Linguistique et Informatique — École Normale Supérieure de Fontenay St Cloud
⋆ Laboratoire d'Informatique de Paris-Nord — Université Paris 13
{pz,jb}@biomath.jussieu.fr, habert@msh-paris.fr, nazarenk@ura1507.univ-paris13.fr

Résumé

Même dans des domaines très spécialisés comme la médecine, il s'avère impossible pour des tâches de traitement automatique basées sur les connaissances de reprendre telles quelles les ontologies existantes. Nous avons examiné l'hypothèse selon laquelle une exploitation raisonnée et systématique d'un corpus représentatif du domaine peut aider soit à opérer un premier repérage des entités et des relations qui caractérisent le domaine soit à modifier ces ontologies.

Nous avons donc effectué une expérience avec ZELLIG un outil d'analyse de corpus. ZELLIG opère un rapprochement d'entrées lexicales sur la base de contextes syntaxiques normalisés. Sur la base d'un corpus constitué dans le domaine des maladies coronaires, nous montrons comment ces rapprochements permettent de dégager une première catégorisation à « gros grain » puis de la raffiner. Les résultats obtenus sont cohérents avec un modèle du domaine, tel qu'on le trouve dans une catégorisation médicale existante, la SNOMED internationale. Ce modèle donne par ailleurs un cadre d'interprétation à la catégorisation émergente, rendant complémentaires les deux approches : analyse de corpus et recours à des connaissances du domaine.

Un recours méthodique au corpus

Une « ontologie » forme un système conceptuel rendant compte d'un domaine de connaissance, système constitué des catégories d'objets sensés exister dans ce domaine. Le donné ontologique est un préalable à toute modélisation formelle (Gruber, 1995 ; Woods, 1991 ; Bachimont, 1996). Même dans des domaines très spécialisés comme la médecine, il s'avère impossible pour des tâches de traitement automatique basées sur les connaissances de reprendre telles quelles les ontologies existantes (Charlet *et al.*, 1994 ; Zweigenbaum *et al.*, 1995). Force est d'utiliser des corpus de textes du domaine soit pour opérer un premier repérage des entités et des relations qui caractérisent le domaine soit pour modifier ces ontologies. On souhaite alors disposer d'outils évolués (au delà d'un simple concordancier) pour une exploitation raisonnée et systématique d'un corpus représentatif du domaine à modéliser.

ZELLIG (Habert *et al.*, 1996 ; Habert & Nazarenko, 1996) est un outil d'analyse de corpus s'inscrivant dans la lignée des travaux de Zellig Harris (Harris *et al.*, 1989). L'hypothèse est qu'il est possible de mettre en évidence, à partir d'une analyse distributionnelle de contextes rendus élémentaires, les classes de concepts et de relations d'un sous-langage lié à un secteur d'activité.

Partant d'une double expérience de modélisation conceptuelle et d'acquisition sur des bases linguistiques de catégories propres à un corpus (Zweigenbaum *et al.*, 1995 ; Bouaud *et al.*, 1995 ; Habert & Nazarenko, 1996 ; Bouaud *et al.*, 1997), notre objectif est ici d'examiner la capacité d'une modélisation linguistique à identifier et structurer des catégories ontologiques. L'expérience est conduite sur le corpus rassemblé pour le projet européen MENELAS (Zweigenbaum & Consortium MENELAS, 1994) de réalisation d'un système de compréhension de comptes rendus d'hospitalisation dans le domaine des maladies coronariennes.

Nous présentons d'abord la méthode de rapprochement d'entrées lexicales sur la base de contextes syntaxiques normalisés. Nous montrons ensuite comment ces rapprochements permettent de dégager une catégorisation à « gros grain », que nous confrontons à une catégorisation médicale existante. Nous examinons également comment ces catégories peuvent être structurées. Enfin, nous soulignons la complémentarité d'une analyse de corpus et du recours à des connaissances du domaine pour une première identification et organisation de catégories ontologiques.

Méthodologie linguistique d'apprentissage de classes

Sur la base d'un corpus, ici celui de MENELAS composé principalement de comptes rendus d'hospitalisation et de lettres de sortie, on cherche à effectuer des regroupements linguistiques constitués de lemmes partageant des comportements linguistiques semblables. Ainsi, 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 (Habert *et al.*, 1996) s'appuie sur les arbres d'analyse produits par des extracteurs de groupes nominaux, ici Lexter (Bourigault, 1994) et AlethIPGN [2], pour en extraire les arbres élémentaires qui y figurent. Par exemple, pour la séquence *stenose severe de le tronc commun gauche* [3], il fournit quatre arbres élémentaires à partir de l'arbre d'analyse complet (voir la figure 1).

Sont considérés comme élémentaires les arbres mettant en évidence une relation binaire entre deux mots pleins, nom ou adjectif, dans des schémas comme, par exemple, `N Prep N` ou `N Adj`. Les dépendances élémentaires ainsi définies n'ont pas forcément de réalisation effective dans le corpus [4] mais ils correspondent à des relations de dépendance vérifiées dans les arbres d'analyse, si l'on passe par une représentation logique de ces arbres et de ces dépendances élémentaires (Gaussier & Habert, 1997).

On obtient les classes de mots possibles dans une position donnée d'une dépendance élémentaire, par exemple les mots qui commutent avec tronc dans l'arbre *b* (figure. 1) :

⋆. À paraître aux Journées Scientifiques et Techniques, Francil, Avignon, 15-16 avril 1997.

2. Développé par GSI-ERLI dans le cadre du projet Eureka GRAAL.

3. 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.

4. Par exemple, on ne trouve pas *stenose de tronc*, dans le corpus, et les arbres *b* et *d* de la figure 1 ne figurent pas directement dans l'arbre complet.

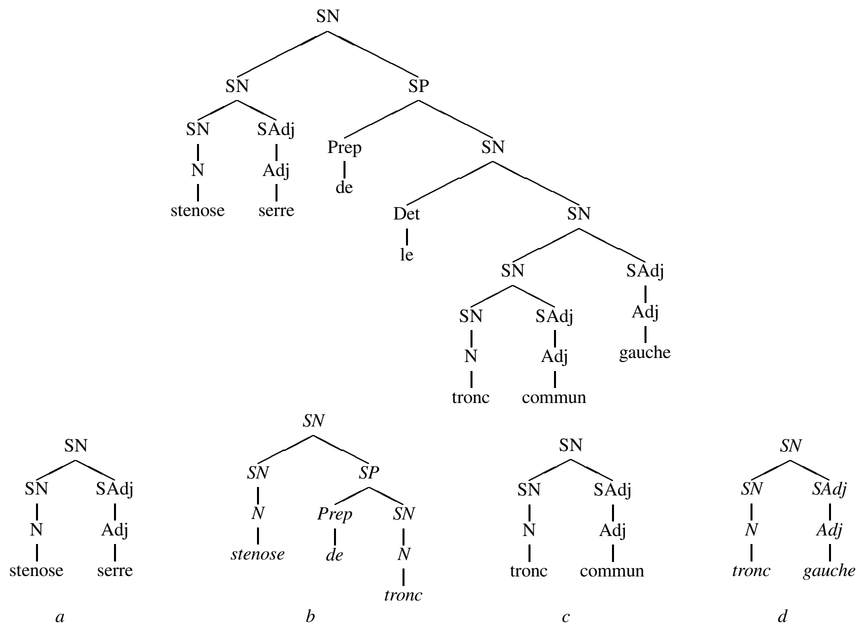

Figure 1 : Arbre d'analyse complet pour *stenose severe de le tronc commun gauche* et arbres élémentaires extraits.

allure, artere (10 o.[5]), branche (3 o.), carotide, debut, diagonale (3 o.), droite (4 o.)... On peut ensuite rapprocher les formes qui partagent des contextes élémentaires (dans l'exemple : artere, carotide, branche, segment entrent dans un syntagme `N Prep N` où le premier `N` est stenose).

ZELLIG, à partir des contextes élémentaires, construit un graphe dont les lemmes sont les nœuds et les contextes partagés par deux lemmes les arêtes. Un nombre minimal de contextes partagés est requis pour placer une arête entre deux lemmes : dans cette expérience, soit 5, soit 10. Ce graphe fournit deux formes de regroupement de lemmes à travers ses composantes connexes (CC) et ses cliques (KC)[6]. La figure 2 montre la septième CC obtenue pour AlethIPGN au seuil 10 (CCA10-7[7]). On notera sur chaque arête les contextes partagés par les lemmes-sommets.

Au total sur le corpus MENELAS, on obtient 30 CC (13 au seuil 5 et 17 au seuil 10) et 92 KC, toutes au seuil 10. Le tableau 1 indique le volume couvert par les syntagmes nominaux extraits par AlethIP et Lexter, en nombre de mots pleins (nom, adjectif, participe passé, participe présent, adverbe, inconnu (nom), prefixe, sigle), ainsi que celui des CC trouvées par Zellig.

Catégorisation à gros grain

Dans un premier temps, un examen des regroupements identifiés par ZELLIG cherche, selon l'hypothèse de Harris (1989), à dégager une première catégorisation, même « grossière », du domaine. Ces regroupements sont ensuite confrontés à une catégorisation médicale existante, la nomenclature « SNOMED internationale » (Côté *et al.*, 1993).

5. o. = occurrences.

6. 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.

7. 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.

Corpus total (tous mots confondus, non lemmatisé)						
Formes ≠	6191					
Occurrences	84839					
Sous-corpus SN (mots pleins uniquement, lemmatisé)						
	AlethIP		Lexter		Union	
Mots ≠	3163		3032		3683	
Occ.	23727		23124		32652	
CC						
Seuil	5	10	5	10	5	10
CC	5	10	8	7	13	17
Mots ≠	250	77	147	33	261	79
%	7,9	2,4	4,8	1,0	7,0	2,1
Occ.	12273	6485	9454	4279	12375	6746
%	51	27	40	18	37	20

Tableau 1 : Volume couvert (mots pleins).

Examen des regroupements linguistiques

Sur les données produites par ZELLIG, quels que soient l'extracteur et le seuil utilisés, on observe une répartition similaire des résultats : une ou deux CC pléthorique(s) qui semble(nt) regrouper plusieurs catégories sémantiques en intersection, une série ensuite de CC beaucoup plus petites (parfois réduites à 2 ou 3 nœuds) et plus homogènes.

C'est le cas de CC produites par ZELLIG sur les sorties de Lexter au seuil 5[8], où les CC 1 et 2 s'opposent aux autres :

CCL5-1 : (99 nœuds, 213 arêtes) { test epreuve examen therapeutique effort angine gene partie dyskinesie pas existence absence aspect presence recidive syndrome myocarde dyspnee anevrisme calibre aval carotide interventriculaire droite hospitalisation intervention revascularisation calcification occlusion pathologie onde dysfonction trouble elevation surcharge hypertrophie

8. Les lemmes et leurs contextes partagés forment ici un graphe de 1909 nœuds et 30988 arêtes, avec un maximum de 309 contextes partagés entre deux lemmes.

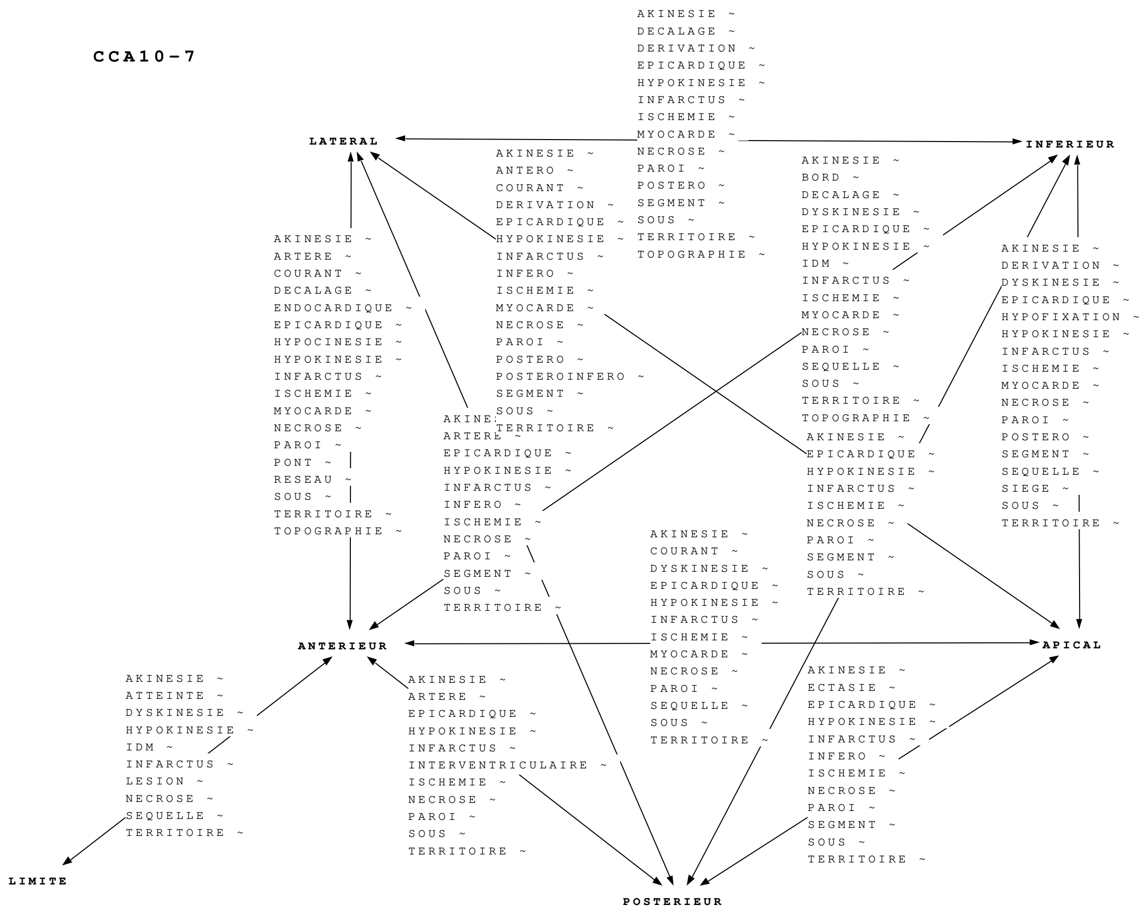

Figure 2 : La composante connexe obtenue pour AlethIPGN au seuil 10 (CCA10-7).

modification alteration souffle instabilite etat contractilite cinetique fraction dilatation pression fonction cavite ventricule marginal diagonale dernier 3eme 2eme premier tronc pont branche reseau plaque obstruction retrecissement maladie atherome lesion spasme stenose angioplastie pontage artere angiographie atteinte anomalie symptomatologie evolution signe manifestation angor sequelle derivation territoire ischemie paroi infarctus akinesie segment necrose insuffisance arythmie douleur dose traitement reapparition controle exploration plan bilan mois jour }

CCL5-2 : (34 nœuds, 64 arêtes) : { minime circonflexe arteriel pulmonaire cardiaque coronaire antero-lateral antero-apical apical lateral posterieur postero-inferieur inferieur normal gauche ventriculaire myocardique anterieur droit important valvulaire mitral calcifie aortique diffus severe ancien recent actuel coronarien proximal distale moyen significatif }

CCL5-3 : (3 nœuds, 2 arêtes) { fonctionnel clinique electrique }

CCL5-4 : { discret net }

CCL5-5 : { court long }

CCL5-6 : { realisation indication }

CCL5-7 : { mauvais beau bon }

CCL5-8 : { petit gros }

Nous allons examiner les regroupements possibles au sein de CC volumineuses sur l'exemple de la CCL5-2 que montre la figure 3 [9]. Comme on le voit par comparaison avec la liste plate donnée ci-dessus, la simple visualisation du graphe permet déjà de percevoir des regroupements à travers la densité relative des liens entre les différents nœuds [10]. En effet, ces CC mettent en évidence des lemmes « attracteurs », qui partagent des contextes avec de nombreux autres lemmes. C'est le cas ici pour coronarien, anterieur, important. Autour de ces attracteurs, on remarque des zones denses, enchevêtrées, où les cliques foisonnent. Dans cet exemple, { anterieur lateral inferieur posterieur }, { coronarien coronaire gauche droite }, { important minime severe significatif }. Enfin, des groupes de nœuds dont le rattachement au reste de la CC est ténu. L'arête coronarien-proximal représente le seul lien à la CCL5-2 de l'étoile autour de proximal : { proximal distale moyen }. Cependant, certains nœuds (recent ancien actuel normal calcifie) ne trouvent pas leur place dans ce premier découpage qui oppose essentiellement des localisants à des évaluatifs.

L'interprétation des CC déclenche deux types de réactions. La volonté de voir si certains regroupements, a priori surprenants, ne manifestent pas toutefois des comportements sémantiques propres au domaine. C'est le cas de certaines

9. Pour des raisons de lisibilité, les arêtes produites par ZELLIG ne sont pas portées sur cette figure.

10. D'autant que l'outil de visualisation que nous utilisons (GraphX ftp://mycroft.ens-fcl.fr/pub/graphx/), développé par Serge Heiden (ELI), préserve la connectivité du graphe lorsqu'on déplace les nœuds, ce qui permet d'essayer les configurations qui maximisent les proximités. La transformation du placement initial des nœuds est alors déjà une interprétation.

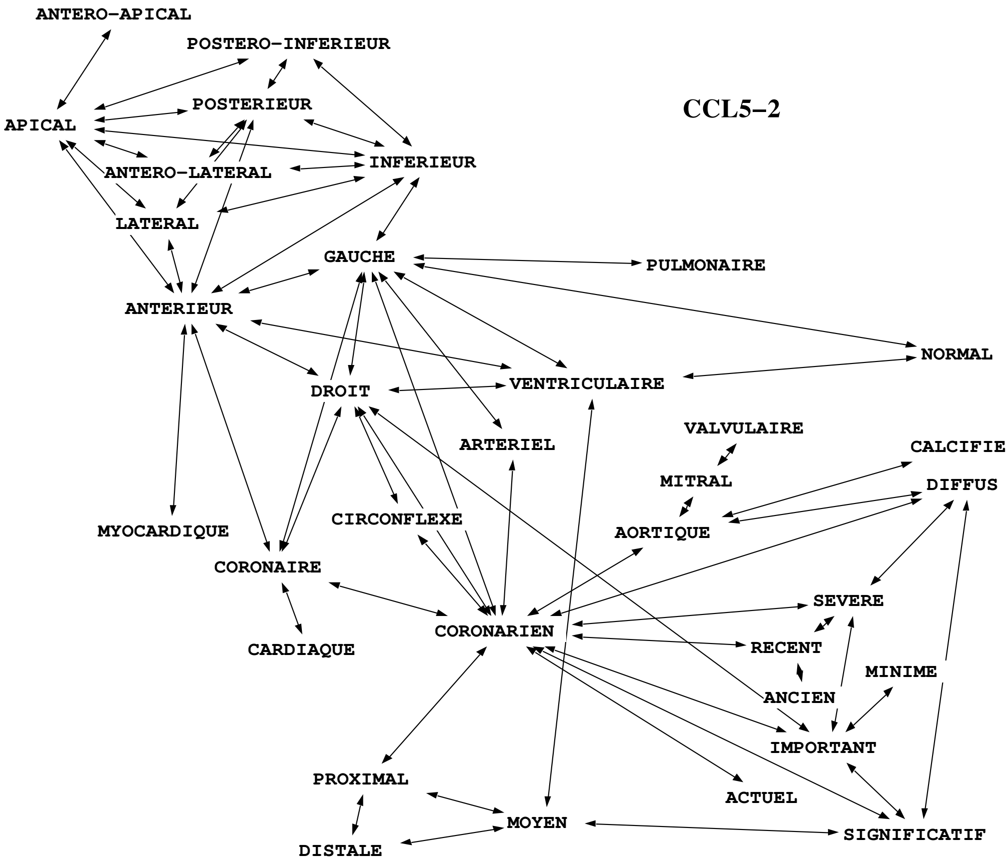

Figure 3 : La deuxième composante connexe obtenue pour Lexter au seuil 5 (CCL5-2).

CC à trois nœuds où un nœud joue le rôle d'intermédiaire, par exemple mauvais ⟵ (∼ [11] *aval, ∼ etat, ∼ fonction, ∼ pronostic, ∼ qualite, ∼ rapport, ∼ resultat, ∼ ventricule*) ⟶ bon ⟵ (∼ *aval, ∼ branche, ∼ calibre, ∼ lit, ∼ qualite*) ⟶ beau. Dans ce type de texte, beau fonctionne comme un diagnostic spécifique sur les possibilités d'évolution d'un site corporel. Hors connaissance du domaine, on ne saurait prédire les mots qu'il modifie effectivement. Ce qui semble 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, ou qui facilitera la réorganisation du réseau coronaire après intervention chirurgicale. Seconde réaction : la volonté de faire « éclater » certains regroupements qui paraissent hétérogènes, en fonction de connaissances générales sur la langue ou sur le monde et au regard des contextes élémentaires qui annotent les arêtes. Ou encore de subdiviser certains catégories encore trop grossières, comme les « localisants », évoquées ci-dessus.

Construction des catégories du domaine

Pour examiner la validité de ces premiers regroupements sur une base syntaxique, nous avons substitué aux lemmes sommets d'une CC/KC une catégorie conceptuelle issue de la la nomenclature « SNOMED internationale » (Côté *et al.*, 1993), nomenclature très employée dans le domaine médical. Nous avons employé les 11 catégories de plus haut niveau de la SNOMED : Topographie (T), Morphologie (M), Fonction (F), Organismes vivants (L), Médicaments, produits chimiques et biologiques (C), Agents, activités physiques et forces naturelles (A), Métiers et professions (J), Contexte social (S), Maladies et diagnostics (D), Procédures et actes professionnels (P), Qualificatifs et termes relationnels (G). Nous avons ainsi catégorisé 937 des 994 lemmes du lexique sémantique de MENELAS [12]. Le tableau 2 indique le nombre de lemmes par catégorie SNOMED, et le nombre d'occurrences correspondant selon AlethIP et Lexter.

Cat	Lemmes/Occurrences, selon			
	AlethIP		Lexter	
A	19	320	16	299
D	22	271	23	350
G	184	4105	188	3680
M	48	1250	48	1223
S	2	30	2	22
C	33	177	41	266
F	56	1600	57	1753
L	8	531	8	516
P	53	1357	54	1675
T	60	2418	61	2680
total				
10	485	12059	498	12464

Tableau 2 : Catégories « SNOMED ».

11. Les deux lemmes du contexte partagé, ici mauvais et bon, peuvent figurer à la place de ∼.

12. Seule une partie de la SNOMED est actuellement traduite en français : le répertoire d'anatomopathologie (12500 termes) (Côté, 1996), que nous remercions le Docteur Roger A. Côté d'avoir gracieusement mis à notre disposition.

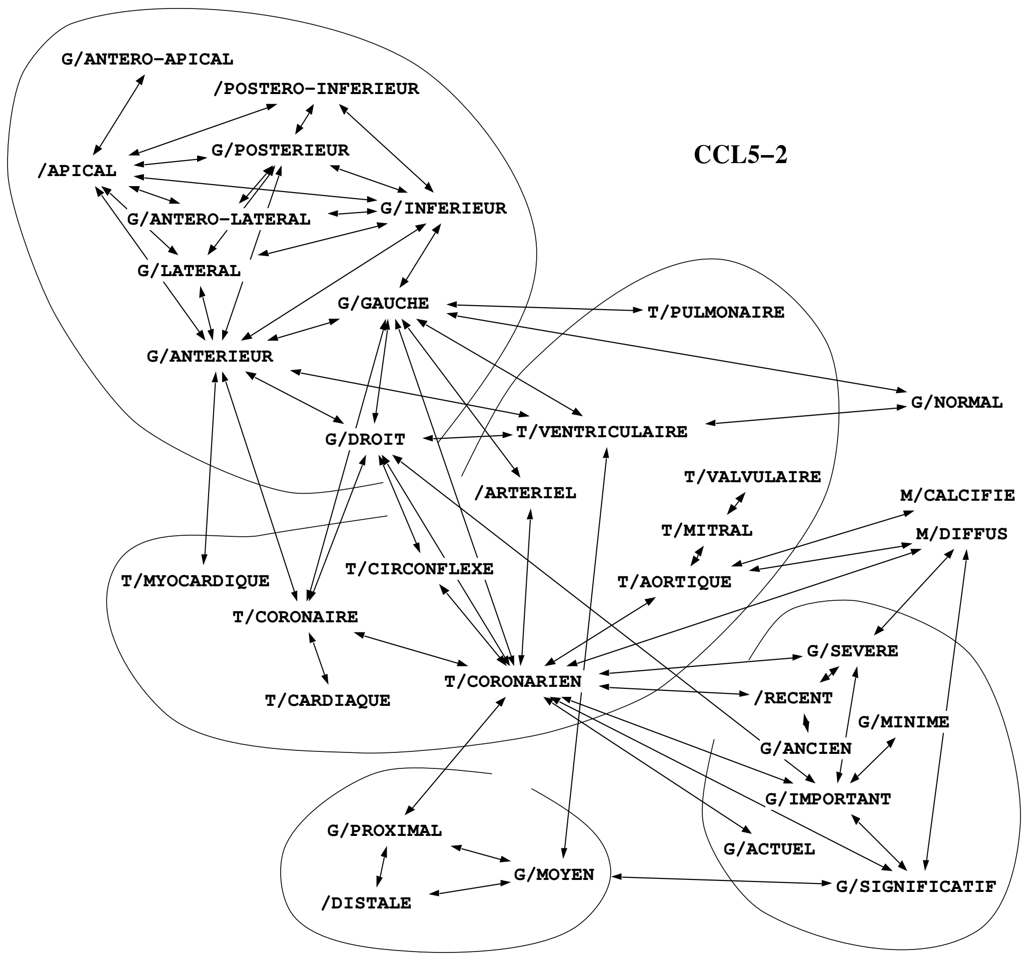

Figure 4 : Regroupements de lemmes de même catégorie SNOMED dans CCL5-2. Les lemmes catégorisés sont précédés d'une étiquette (ex. T/MITRAL), les autres n'en comportent pas (ex. /ARTERIEL).

Examen de l'homogénéité des CC. L'espoir que nous avions était qu'une CC regroupe des lemmes appartenant à une catégorie sémantique commune, ou au pire à des catégories sémantiques très proches. Il a été largement confirmé. Si l'on considère l'ensemble des 30 CC trouvées, on observe que 20 sont homogènes, 2 contiennent 2 classes très proches, et 2 étaient inhomogènes, mais du fait d'incohérences dans l'étiquetage. Au total donc, 24 CC sur 30 confirment notre attente.
Parmi les 6 restantes, une contenait un « intrus » :

{$_{\mathrm{CCL}10\text{-}2}$ effort/F douleur/F angor/D}

angor est catégorisé comme un diagnostic, alors que les deux autres lemmes sont étiquetés comme des « fonctions » (pathologiques ou physiologiques). De fait, dans le lexique de MENELAS, angor était ambigu, mais cette ambiguïté n'a pas été reportée dans la catégorisation SNOMED.
Les 5 autres CC sont grosses à très grosses (14, 34, 50, 99, 233 lemmes). Nous avons essayé d'exploiter leur structure graphique, en tenant compte de la connectivité des nœuds. Nous supposons que les nœuds d'une même catégorie sémantique se regroupent dans des sous-graphes connexes. Lorsqu'on applique ce principe à CCL5-2, on obtient les regroupements suivants (figure 4) :

G (modifieurs) : {gauche droit anterieur inferieur lateral antero-lateral posterieur apical postero-inferieur antero-apical} {moyen proximal distale}
cf aussi {$_{\mathrm{CCA}10\text{-}7}$ posterieur anterieur apical inferieur lateral} ;
{important significatif severe minime recent ancien}
cf aussi {$_{\mathrm{CCA}10\text{-}5}$ modere net important minime significatif}

T (topographie) : {coronaire coronarien cardiaque circonflexe aortique mitral valvulaire} *cf aussi* {$_{\mathrm{CCA}10\text{-}3}$ coronarien coronaire} *cf aussi* {$_{\mathrm{CCA}10\text{-}8}$ aortique mitral}

De plus, on constate que ces sous-graphes sont souvent relativement disjoints du reste du graphe (voir la figure 4).
En résumé, les CC produites par ZELLIG créent des regroupements de lemmes dont la comparaison avec le premier niveau de la SNOMED montre la pertinence ontologique. Si l'on élimine la plus grosse CC[13], les 29 CC restantes permettent de déterminer 37 groupes homogènes dont plusieurs se recoupent. Certains d'entre eux correspondent à des synonymies, comme {$_{\mathrm{CCA}10\text{-}3}$ coronarien coronaire}. Notons que nous n'avons pas ici examiné les catégories SNOMED des *arêtes*, qui peuvent apporter des informations supplémentaires.

Heuristique de catégorisation après amorçage. Comme indiqué ci-dessus, les lemmes étiquetés souvrent seulement une partie du corpus examiné. De ce fait, certains lemmes des CC ne sont pas étiquetés. Nous avons examiné dans

13. Celle-ci était trop volumineuse pour être chargée sous GraphX.

quelle mesure le travail de ZELLIG pourrait aider à catégoriser ces lemmes sur la base des lemmes déjà étiquetés. L'idée est de se fonder sur l'hypothèse formulée ci-dessus, selon laquelle les lemmes de même catégorie sont regroupés dans les CC. Nous nous appuyons sur cette hypothèse pour formuler deux heuristiques d'étiquetage, dont la seconde généralise la première.

Heuristique 1 (Étiquetage, CC homogène) *Les lemmes non étiquetés d'une CC homogène prennent la catégorie sémantique du reste de la CC.*

Cette heuristique s'applique, parmi les 25 CC considérées comme homogènes (toutes sauf les 5 « grosses »), aux 16 qui contiennent un ou plusieurs lemmes sans étiquette. Elle permet ainsi d'aider à étiqueter 25 lemmes sans étiquette. Elle révèle aussi un étiquetage incohérent pour trois lemmes dans deux groupes : on peut se demander si, par exemple, $\{_{\text{CCL5-5}}$ long/M court/G$\}$ ne devraient pas être soit tous deux G (qualificatif), soit tous deux M (morphologie) ; leur considération séparée dans l'étiquetage des lemmes a probablement conduit à une incohérence.

Heuristique 2 (Étiquetage, CC non homogène) *Soit un lemme non étiqueté d'une CC non homogène. Sa catégorie sémantique est choisie à la majorité absolue de celles de ses voisins.*

On voit que l'heuristique 1 est un cas particulier de celle-ci. Nous avons appliqué l'heuristique 2 aux 4 grosses CC (la plus grosse étant toujours éliminée), ainsi qu'à la petite CC mixte CCL10-3. Par exemple, dans CCL5-2, /apical /postero-inferieur /distal /recent reçoivent correctement la catégorie G (à l'unanimité pour les 3 premiers, à 2 contre 1 pour /recent) ; /arteriel est en ballotage à 1 contre 1 (G/gauche T/coronarien), et n'est donc pas étiqueté par cette heuristique. Ces 5 CC contiennent 46 lemmes non étiquetés, dont 38 sont étiquetables par l'heuristique 2 (pour les 8 autres, une majorité absolue n'est pas obtenue parmi les voisins). Parmi ceux-ci, 26 sont étiquetés correctement, 4 soulèvent un doute requérant un examen du corpus, et 8 constituent des erreurs.

Si l'on exclut également, en parallèle avec CCA5-1, la plus grosse CC de Lexter (CCL5-1), on supprime 5 erreurs et les 4 cas de doute (ainsi que 14 étiquetages corrects). En agrégeant les résultats obtenus sur toutes les CC sauf les deux plus grosses, on a alors 35 étiquetages corrects, 3 corrections d'étiquetages incohérents, et 3 erreurs, couvrant respectivement un nombre d'occurrences de 1439, 165 et 123. Les CC mises en jeu pour aider à produire ces étiquetages contiennent 99 lemmes étiquetés.

Notons bien que ces heuristiques sont considérées comme devant fournir une aide à la personne qui met au point la classification des mots du corpus. Les erreurs sont de ce fait en général facilement identifiées. En tout état de cause, les propositions d'étiquetage devraient faciliter le travail de classification et réduire les incohérences, comme cette expérience l'a illustré.

Structuration des catégories

Si les premiers regroupements opérés sur la base des graphes issus de ZELLIG sont effectivement interprétables en termes de catégories ontologiques, ces premières catégories demandent néanmoins à être affinées et structurées. Là encore, nous étudions cette sous-catégorisation sur des bases linguistiques et à l'aide de connaissances du domaine.

Sur des bases linguistiques

Pour tenter de repérer une sous-catégorisation plus fine, l'examen des arêtes des graphes et des dépendances binaires dans lesquelles entrent les lemmes du corpus est utile. La liste de ces dépendances regroupe l'ensemble des contextes élémentaires d'un lemme et constitue en effet un outil d'exploration textuelle et d'analyse qui s'apparente à une concordance synthétique.

Reprenons ici l'exemple des adjectifs de localisations relatives, étiquetés G. La densité des liens dans certaines zones de CCL5-2 et l'existence de cliques fait apparaître trois groupes : {anterieur lateral inferieur posterieur antero-apical antero-lateral postero-inferieur}, {anterieur gauche droit ventriculaire coronaire coronarien} et {proximal moyen distal}. L'examen des contextes qui rapprochent ces différentes formes permet effectivement de cerner les contours de trois sous-classes :

1. Localisants du myocarde : anterieur, posterieur, apical, lateral... La proximité de ces adjectifs, déjà apparente sur la figure 3 se trouve confirmée par la figure 2, issue d'une autre expérience. Dans cette dernière, les contextes territoire ∼ et topographie ∼ indiquent clairement qu'il s'agit de localisants et on constate que ces adjectifs modifient essentiellement des noms désignant le coeur ou ses parties (myocarde, endocardique, epicardique) ou des phénomènes cardiaques (infarctus, decalage, ischemie, hypokinesie, akinesie...). L'appartenance à ce groupe de anterieur, qui fait la frontière avec le reste de la CC dans la figure 3 est en fait sans ambiguïté : près de la moitié de ses contextes sont explicitement « cardiaques », les autres étant des contextes plus génériques (reseau ∼, sequelle ∼, atteinte ∼...). En revanche, la proximité avec limite est moins nette : ce lemme est présent dans CCA10-7, mais marginalisé ; s'il a plusieurs contextes en commun avec anterieur, ceux-ci sont beaucoup moins fréquents que ceux que anterieur partage avec ses autres voisins.
2. Localisants artériels : proximal, moyen et distal. Comme le montrent les contextes artere ∼, carotide ∼, IVA ∼, branche ∼, segment ∼..., ces adjectifs décrivent les valeurs d'un attribut de localisation qui caractérise la famille sémantique des parties d'artère et par métonymie des artères.
3. Localisants relatifs polyvalents : droit, gauche. Leurs contextes sont plus divers. Ils fonctionnent à la fois comme localisants cardiaques (oreillette ∼, ventricule ∼, lobectomie ∼) et comme localisants artériels (artere ∼, carotide ∼, angiographie ∼). De là, leur place assez centrale dans CCA5-2, entre la famille des localisants cardiaques et celle des noms d'artères identifiées ici par leur nom (interventriculaire, circonflexe...) {arteriel coronarien coronaire...}.

À l'aide de connaissances externes

Les regroupements relativement grossiers sur une base distributionnelle débouchent souvent sur la constitution de sous-catégories plus fines, comme nous l'avons vu pour CCL5-2. 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 autres traitements. Dans la SNOMED, chaque axe (P T etc.) est lui-même subdivisé en sous-classes, jusqu'à 6 niveaux de profondeur. Si l'on examine la hiérarchie des Procédures (P), on trouve effectivement un regroupement des examens (P3–P5), des traitements médicaux (P2), et des opérations chirurgicales (P1) (le troisième groupe sauf hospitalisation).

De façon générale, chaque groupe homogène identifié constitue potentiellement une sous-classe de la catégorie SNOMED commune à ses lemmes. De plus, certains groupes sont inclus dans d'autres. Par exemple, {$_{\mathrm{CCA10\text{-}8}}$ mitral aortique} est inclus dans {$_{\mathrm{CCL5\text{-}2}}$ coronaire coronarien cardiaque circonflexe mitral aortique valvulaire}, et pourrait en constituer une sous-classe : « éléments qui possèdent une valve » ⊂ « éléments du système cardio-vasculaire » ⊂ « parties du corps » (T). De fait, la catégorie SNOMED « appareil circulatoire » (T-30000) regroupe tous les termes indiqués ; et sa sous-catégorie « valve cardiaque » (T-35000) regroupe « valve mitrale » (T-35300) et « valve aortique » (T-35400) [14].

De même, {$_{\mathrm{CCA10\text{-}10}}$ absence pas} (expression de non-existence) est inclus dans {$_{\mathrm{CCL5\text{-}1}}$ existence absence presence aspect pas recidive} (indicateurs de mode d'existence), lui-même inclus dans G (Qualificatifs et termes relationnels). Cette dernière catégorie est celle qui donne lieu au plus grand nombre de sous-classes :

G : {$_{\mathrm{CCL5\text{-}2}}$ gauche droit antero-lateral postero-inferieur antero-apical
{$_{\mathrm{CCA10\text{-}7}}$ posterieur anterieur apical inferieur lateral}}
{$_{\mathrm{CCL5\text{-}2}}$ moyen proximal distale}
{$_{\mathrm{CCL5\text{-}2\ CCA10\text{-}5}}$ important significatif minime
{$_{\mathrm{CCL5\text{-}2}}$ severe recent ancien}
{$_{\mathrm{CCA10\text{-}5}}$ modere net}}
{$_{\mathrm{CCL5\text{-}1}}$ existence presence aspect recidive
{$_{\mathrm{CCA10\text{-}10}}$ absence pas}}
{$_{\mathrm{CCL5\text{-}1}}$ modification elevation alteration trouble}

De fait, alors que les autres axes correspondent à des notions techniques assez bien cernées, l'axe G englobe tout un éventail de modifieurs et de relations, dont le nombre de lemmes et d'occurrences est d'ailleurs largement supérieur à ceux des autres axes (tableau 2). La partie de l'axe G total de la SNOMED incluse dans le répertoire d'anatomopathologie (23 sur 1373) est cependant trop faible pour que nous ayons pu corroborer les sous-classes trouvées avec celles présentes dans la SNOMED. Pour les autres axes, la sous-catégorisation mise en évidence par ZELLIG est cohérente avec celle déjà en place dans la SNOMED.

14. 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.

Deux types de connaissances complémentaires

L'expérience réalisée montre que l'analyse linguistique effectuée par ZELLIG met en évidence des classes et sous-classes de lemmes dont la confrontation à une catégorisation existante, la SNOMED, a montré la pertinence. Assigner aux mots d'un corpus des classes sémantiques hiérarchisées peut être vu comme un préalable à la construction d'une ontologie, pour laquelle il prépare le terrain. Cette tâche se divise en deux parties : le choix des catégories, et l'assignation des mots du corpus aux catégories. Au départ, on se trouve dans une situation dans laquelle on ne dispose pas a priori d'un jeu de catégories complet : c'est justement cette classification que l'on cherche à mettre au point. Une démarche empirique consiste alors à faire des allers-retours entre le choix d'un jeu de catégories temporaire, l'assignation des mots à ces catégories, et le raffinement des catégories courantes pour mieux épouser les propriétés des mots considérés.

Les graphes des contextes partagés font apparaître une première cartographie du domaine. Les regroupements obtenus possèdent une certaine pertinence ontologique. Ils aident à subdiviser certaines catégories. Il faut cependant affiner cette description qui reste grossière et irrégulière. C'est là que l'examen des arêtes des graphes et des dépendances binaires dans lesquelles entrent les lemmes du corpus constitue un précieux outil d'exploration et d'analyse textuelles. Il s'apparente à une concordance synthétique. Il permet de décrire avec une certaine précision les concepts et les attributs qui les caractérisent.

Soulignons pour finir que le seul examen des regroupements et des contextes ne suffit pas à déterminer le jeu des catégories à employer ni leur contour. Il nécessite au contraire l'intervention de connaissances propres au domaine. En outre, les comportements linguistiques attestés s'avèrent insuffisants pour une modélisation conceptuelle : celle-ci doit reconstituer une partie des implicites du domaine qui sont nécessaires aux inférences et à la compréhension. Modélisation conceptuelle et acquisition linguistique de catégories sont donc tout à la fois complémentaires et divergentes.

Références

Bachimont, B. (1996). *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.

Bouaud, J., Bachimont, B., Charlet, J. & Zweigenbaum, P. (1995). Methodological principles for structuring an "ontology". In *Proceedings of the IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing."*, Montréal.

Bouaud, J., Habert, B., Nazarenko, A. & Zweigenbaum, P. (1997). Validité ontologique de catégorisations linguistiques. In *Article soumis à IC'97*.

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

Charlet, J., Bachimont, B., Bouaud, J. & Zweigenbaum, P. (1994). Ontologie et réutilisabilité : expérience et discussion. In *Actes des 5 ^{es} Journées Acquisition des Connaissances*, pp. C1–C14, Strasbourg.

R. A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett & L. Brochu (Eds.) (1993). *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield: College of American Pathologists.

Côté, R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale*. Université de Sherbrooke, Sherbrooke, Québec, v3.4 edition.

Gaussier, E. & Habert, B. (1997). Langue spécialisée : des séquences observées aux mots possibles. In D. Corbin, B. Fradin, B. Habert, F. coise Kerleroux & M. Plénat (Eds.), *Mots possibles et mots existants*, Lille.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, **43**(5/6), 907–928. Special issue on ontology.

Habert, B., Naulleau, E. & Nazarenko, A. (1996). Symbolic word clustering for medium-size corpora. In *16th International Conference on Computational Linguistics*, volume 1, pp. 490–495, Copenhagen, Denmark.

Habert, B. & Nazarenko, A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. In *Journées sur l'acquisition des connaissances*, pp. 137–142, Sète: AFIA.

Harris, Z., Gottfried, M., Ryckman, T., Mattick Jr, P., Daladier, A., Harris, T. & Harris, S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science*. Dordrecht, The Netherlands: Kluwer Academic Publisher.

Woods, W. A. (1991). Understanding subsumption and taxonomy: A framework for progress. In J. F. Sowa (Ed.), *Principles of Semantic Networks*, chapter 1, pp. 45–94. San Mateo, Ca.: Morgan Kaufmann Publishers.

Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J. & Boisvieux, J.-F. (1995). Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of Information in Medicine*, **34**(1/2).

Zweigenbaum, P. & Consortium MENELAS (1994). MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, **45**, 117–120.