Corpus-Based Identification and Refinement of Semantic Classes

A. Nazarenko, Ph.D.⋆, P. Zweigenbaum, Ph.D.†, J. Bouaud, Ph.D.†, B. Habert, Ph.D.‡

⋆ Laboratoire d'Informatique de Paris-Nord — Université Paris 13

† DIAM — Service d'Informatique Médicale/AP-HP & Dépt. Biomathématiques U. Paris 6

‡ Équipe Linguistique et Informatique — École Normale Supérieure de Fontenay St Cloud

*Medical Language Processing (*MLP*), especially in specific domains, requires fine-grained semantic lexica. We examine whether robust natural language processing tools used on a representative corpus of a domain help in building and refining a semantic categorization. We test this hypothesis with* ZELLIG*, a corpus analysis tool. The first clusters we obtain are consistent with a model of the domain, as found in the* SNOMED *nomenclature. They correspond to coarse-grained semantic categories, but isolate as well lexical idiosyncrasies belonging to the clinical sub-language. Moreover, they help categorize additional words.*

INTRODUCTION

Medical vocabularies are a fundamental resource for medical information processing.[1] They are faced with a difficult problem of coverage,[2] both in width, with different disciplines and new terms, and in depth, to produce more precise descriptions with modifiers and context. A promising way to extend vocabulary coverage is to examine medical corpora, such as patient narratives, with the help of robust natural language processing tools.[3] This can help propose new terms or modifiers for inclusion in existing vocabularies. An issue then arises of categorizing these new items. We aim to assess the relevance of advanced corpus linguistic tools to identify and structure semantic categories.

ZELLIG[4] is such a tool. It has been designed to automate the discovery of semantic classes in the spirit of Harris' work.[5] Harris claims it is possible, with a distributional analysis of elementary contexts, to isolate the concepts and the relationships of the sub-language of a given domain. We ran an experiment with ZELLIG on the corpus gathered for the European MLP project MENELAS[6] in the domain of coronary diseases.

Much research has focused on the automatic construction of semantic classes from corpora. General lexical databases like WordNet[7] or *Roget's thesaurus* do not describe the technical and specialized word uses, and hand-crafting specialized terminologies and thesauri corresponds to long-run tasks. The main approaches to build specialized semantic classes consist either in specializing, *i.e.*, contextualizing, general semantic relations,[8] or in acquiring specific semantic relations from the distributions of words.[9,10]

The present work belongs to the second approach. Our work aims at extracting not only similarity relationships[9] between words or even semantic axes[10] but also to group words into classes refering to coherent semantic categories. In that respect, our objective is related to that of Bensch and Savitch.[11] However, they rely on an automatic classification algorithm whereas we consider that interpretation is central in the categorization process. We present a corpus exploration method to help that interpretation process. This method relies both on extracting semantic information from the corpus data and projecting semantic knowledge in a manner close to Basili et al.[12]

We first present the methodology for grouping words relying on normalized syntactic contexts. We show how these clusters allow us to obtain a coarse-grained categorization. We assess the relevance of this first result by comparing it with the SNOMED international nomenclature.[13] The various ZELLIG labelled graphs help in structuring and refining the first categorization. Last, we discuss the interaction between corpus analysis and domain knowledge in order to build or modify semantic lexica.

A LINGUISTIC METHODOLOGY FOR ONTOLOGICAL CLASS DISCOVERY

Grefenstette[10] distinguishes three types of semantic affinities between words and three steps in discovering semantic categories: *"First-order techniques examine the local context of a word attempting to discover what can co-occur with that word within that context. Second-order techniques derive a context for each term and compare these contexts to discover similar words or terms. Third-order techniques compare lists of similar words or terms and group them along semantic axes"*. Our method follows this three-step process to discover how words can be grouped together within a given domain according to the contexts they share.

ZELLIG uses normalized syntactic noun phrases (NPs) as local contexts for the first-order step. It uses parse trees retrieved by two NP extractors: AlethIPGN (developed within the European Eureka GRAAL project) and Lexter.[14] NPs are generally assumed to express the main notions of a domain, and they do cover a large part of the corpus (between 27 and 38 %, see table 1). ZELLIG automatically reduces these numerous and complex NPs to elementary dependency trees, which more readily exhibit the fundamental binary relations between content words. For instance, from the parse tree for *"sténose serrée du tronc commun gauche"* (*tight stenosis of left main stem*), ZELLIG yields the set of elementary trees of figure 1.

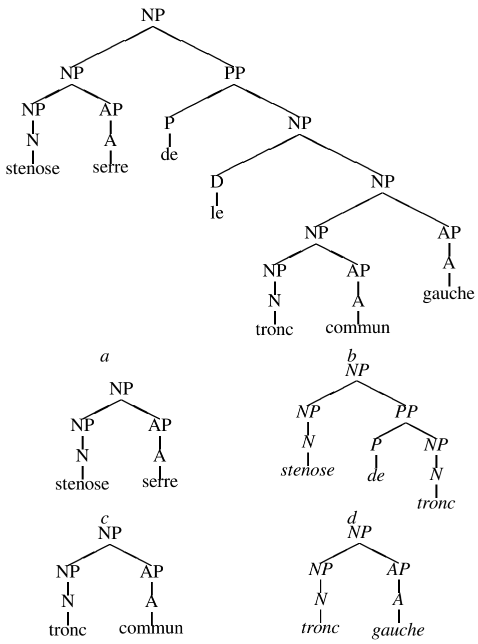

Figure 1: Parse tree and elementary dependencies.

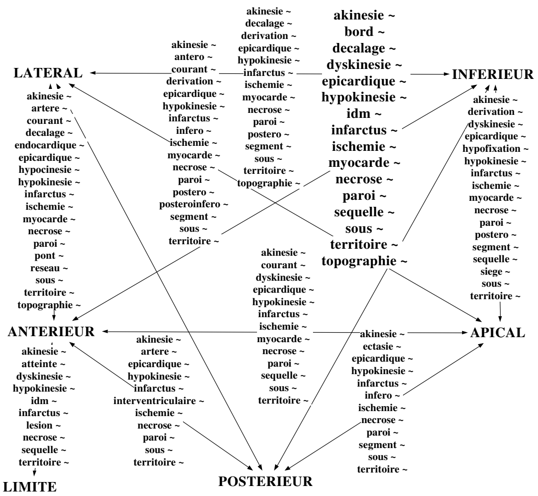

Figure 2: A connected component (CCA10-7).

Second-order affinities show which words share the same environments. For instance, the following words can replace *tronc* in tree *b*: allure, artere (10 occurrences), branche (3 o.), carotide, debut, diagonale (3 o.), droite (4 o.)... That is, they can appear in the same environment: a N P N tree, whose first noun is stenose.

As a third-order technique aiming at deriving subgroupings of similarity, a graph is computed by ZELLIG. The words constitute the nodes. A link corresponds to a certain number of shared contexts according to a chosen threshold (5 or 10 in the experiment). Edges are labelled with the shared contexts. ZELLIG also computes the strongly connected components (CCs: the sub-graphs in which there is a path between every pair of distinct nodes) and the k-cliques (KCs: the sub-graphs in which there is an edge between each node and *every other node* of the graph). These are the most relevant parts of the graph on topological grounds. Figure 2 shows such a CC. On

Table 1: Corpus coverage (content words).

Total corpus (all words, unlemmatized)						
≠ forms	6191					
occurrences	84839					
NP sub-corpus (content words only, lemmatized)						
	AlethIPGN		Lexter		Union	
≠ lemmas	3163		3032		3683	
occ.	23727		23124		32652	
Connected Components						
threshold	5	10	5	10	5	10
# CC	5	10	8	7	13	17
≠ lemmas	250	77	147	33	261	79
% NP lem.	7.9	2.4	4.8	1.0	7.0	2.1
occ.	12273	6485	9454	4279	12375	6746
% NP occ.	51	27	40	18	37	20

the MENELAS corpus, ZELLIG produced 30 CCs (13 at threshold 5 and 17 at threshold 10) and 92 KCs (threshold = 10) (see table 1).

CLUSTERING

Linguistic clustering

The various KCs and CCs that ZELLIG produces were found to belong to two sharply different kinds. Most of them group between two and six lexical entries (*e.g.*, CCA10-7, figure 2). Within such limited sets, the words belong to the same global semantic axis. As such, they help give a semantic tag to the whole set or to some of its nodes according to the semantic label of the others. They are organized by synonymy, antonymy or scalar relationships. These sets also help discover idiosyncratic similarities or oppositions, which are important to build lexically tuned MLP systems, as clinicians happen to give new meanings to ordinary words. For instance, CCL5-2 gathers {mauvais bon beau} (literally, {bad good fine}), which is rather surprising, as there is a discrepancy between the first two adjectives and the last one: the evaluation criteria differ. However, as the shared contexts prove it, in medical records, beau is used as a synonym of bon (literally, good, meaning *fine*). More precisely, it qualifies (parts of) organs (*e.g.*, artery branch) whose overall state is satisfactory. As is obvious in this example, edge labels permit to check immediately the possible semantic categorizations for the nodes of the graphs, or to isolate the odd ones (such as limite (*limit*) in figure 2).

CCL5-2 belongs to the second kind of CCs, which have many more nodes (34, 50, 99, 233). It is shown on figure 3 (without edge labels; nodes have been additionally tagged, as explained later). Here, the observation of graph topology permits to isolate consistent sub-areas, as the set {lateral apical anterieur posterieur inferieur}, which by the way constitutes a KC. As for the limited KCs or CCs discussed above, as soon as these groups have been pointed out, their immediate neighbors can be assimilated to the same semantic category: this heuristic can be applied to {antero-apical postero-inferieur antero-lateral} which use components of the category core. Some words act as mediators between groups. That is the case for gauche (*left*) which is articulated with the first semantic category as well as with a second one {droite coronaire circonflexe arteriel coronarien}. Such a situation can give rise to the hypothesis of shades of meaning for a given word, or even of the existence of homonyms. The pair {coronaire coronarien} presents such shades of meaning: the former is more descriptive, the latter more evaluative, as it is associated to adjectives such as {severe important significatif}, which constitute a third semantic category.

The visual map of relationships between words provided by ZELLIG enables, at first glance, to identify coarse-grained semantic categories and to isolate lexical idiosyncrasies belonging to the clinical sublanguage. However, the conceptual validity of these linguistic groupings must be assessed.

Cluster validation

We therefore decided to confront them to an existing categorization: the SNOMED International nomenclature,[13] which both has very good clinical coverage[2] and is partially translated into French. We started from the 11 high-level SNOMED categories (T M F L C J A S D P G). We first categorized 937 out of the 994 lemmas of the MENELAS semantic lexicon. This was mainly performed manually, since we worked on simple lemmas rather than multi-word terms, and because only part of the SNOMED terms were available to us in French: the Microglossary for Pathology (approx. 12500 terms). We then projected these categories on the lemmas on the CC/KC nodes. Table 2 shows, for the NP sub-corpus, the number of categorized lemmas per SNOMED category and the corresponding number of occurrences according to AlethIPGN and Lexter.

Our hope that a CC would group together lemmas that belong to a common semantic category was widely confirmed. We examined the 30 CCs, and found that 20 were homogeneous. Two were inhomogeneous, but probably due to tagging inconsistencies. One can wonder, for instance, whether {long/M court/G} (*short*) should not be either both G (modifier) or both M (morphology). Their separate consideration when tagging the lexicon probably lead to an inconsistency. In total then, 22 CC out of 30 fulfilled our expectations.

Table 2: Occurrences of SNOMED-tagged lemmas.

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

Three of the remaining 8 contained one outsider, *e.g.*, {effort/F douleur/F angor/D}: we categorized angor (*angina*) as a diagnosis, whereas the two other lemmas were tagged as "functions". The other 5 CCs are the largest ones. One can observe though that nodes with the same semantic category cluster together in connected sub-graphs. For instance, in CCL5-2, one can find the following clusters (figure 3); moreover, these sub-graphs often happen to be relatively disjoint from the rest of the graph:

G (modifiers) : {gauche droit anterieur inferieur lateral antero-lateral posterieur apical postero-inferieur antero-apical} {moyen proximal distale} {important significatif severe minime recent ancien}

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

In summary, the CCs produced by ZELLIG create lemma clusters which the comparison to the first SNOMED level shows to be relevant. If we eliminate the largest CC (which we could not display), the remaining 29 CCs revealed 37 homogeneous clusters, several of which intersect.

EXTENDING EXISTING CLASSES

As mentioned above, categorized lemmas only cover part of the corpus. Therefore, some lemmas in the CCs remain untagged. We examined the extent to which ZELLIG output could help categorize these lemmas on the basis of already tagged lemmas. We derive from the above observation a tagging heuristic: *given an untagged lemma in a CC, its semantic category is chosen by absolute majority of the ones of its neighbors.* As a trivial case, untagged lemmas in a homogeneous CC get the semantic category of the rest of the CC.

Applying the heuristic to CCL5-2 (figure 3) correctly assigns category G to /apical /postero-inferieur /distale /recent (unanimously for the first 3, 2 against 1 for /recent); /arteriel obtains a tie with 1 against 1 (G/gauche T/coronarien), and therefore does not get tagged. Considering again the untagged lemmas in all CCs but the largest, this heuristic tagged 46 and left 10 untagged.

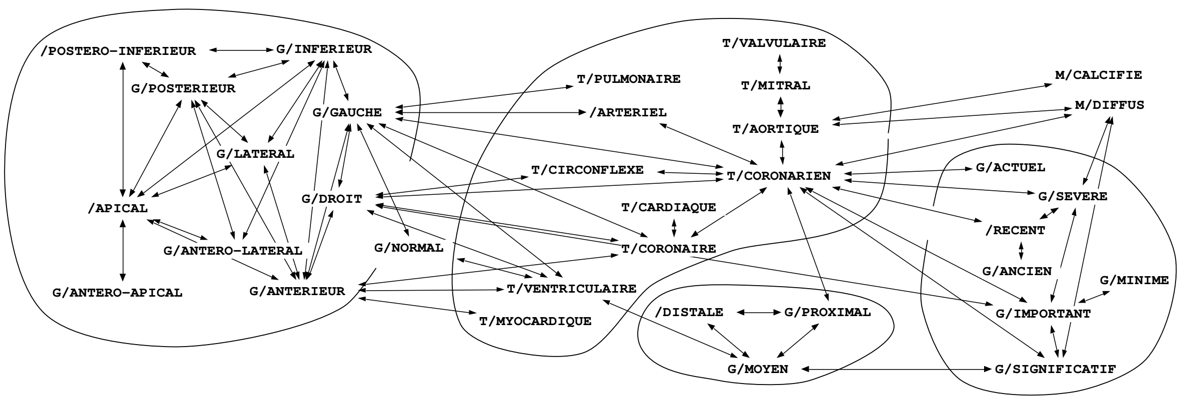

Figure 3: Grouping lemmas with same SNOMED category in CCL5-2. Categorized lemmas are preceded with a tag.

38/46 taggings were correct, 4 were erroneous, and 4 raise a doubt which requires to go back to the corpus. The CCs involved to help obtain these taggings contain 87 tagged lemmas.

REFINING THE HIERARCHY

The Subcategorization Process

ZELLIG graphs show various lemma groups, but the larger clusters must be split and structured. ZELLIG results can help this subcategorization task too. Let us consider the example of the relative localization adjectives. The contrasts of edge density in CCL5-2 and the existence of cliques bring out three different lemma groups: {anterieur lateral inferieur posterieur apical antero-apical antero-lateral postero-inferieur}, {anterieur gauche droit ventriculaire coronaire coronarien} and {proximal moyen distal}. These subclasses can be labelled and their boundaries delimited:

1. Localization relative to the myocardium: anterieur, posterieur, apical, lateral... The proximity between these adjectives is noticeable both in CCL5-2 and CCA10-7. The contexts territoire ∼ and topographie ∼ on the edges of the figure 2 show that the node lemmas are localization adjectives. More specifically, these adjectives modify nouns that refer to the heart and its parts (myocarde, endocardique, epicardique) or to heart phenomena (infarctus, decalage, ischemie, hypokinesie, akinesie...). Although anterieur is a bridge to the rest of the CC (fig. 3), it belongs to that subclass: except for generic ones (reseau ∼, sequelle ∼, atteinte ∼...), most of its contexts are explicitly "cardiac". On the contrary, limite does not belong to that class: in CCA10-7, it is related to this group of lemmas by a single link to anterieur and the frequencies of the shared contexts show that anterieur is much closer to its other neighbors than to limite.

2. Localization relative to arteries: proximal, moyen et distal. As proved by their contexts artere ∼, carotide ∼, IVA ∼, branche ∼, segment ∼..., these adjectives characterize artery parts or arteries.

3. Multi-purpose relative localization: droit, gauche. They occur in more varied contexts. They are used to localize relatively to the heart (oreillette ∼, ventricule ∼, lobectomie ∼) but also relatively to an artery (artere ∼, carotide ∼, angiographie ∼). Hence their central position in CCL5-2, between the heart localization family and the family of artery nouns (interventriculaire, circonflexe, arteriel, coronarien, coronaire...).

Validation of subcategories

In the SNOMED, each category (P T etc.) is itself subdivided into subcategories, up to 6 levels down. As far as we could check, the subcategorization evidenced by ZELLIG is generally consistent with it. CCL5-1, *e.g.*, evidences three clusters of procedures (P), two of which are also found in CCA10-1, giving the subgroups:

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

One can recognize there examinations, (medical) treatments, and invasive treatments (plus hospitalisation), which are found in distinct subcategories (respectively, P3–P5, P2, and P1) of the SNOMED P category.

Where the SNOMED is less developed, as in the G axis (modifiers), ZELLIG could even propose new subcategories. For instance, {absence pas} (non-existence) is included in {existence absence presence aspect pas recidive} (indicators of mode of existence), which should logically belong to G. Modifiers are frequent in medical narratives: among the 20 most frequent lemmas, 4–5 are modifiers according to AlethIPGN–Lexter, and modifier clusters are the most numerous.

DISCUSSION

Organizing the words of a domain in a set of hierarchical categories can be divided into two parts: specifying the categories, and assigning words (found in a representative corpus) to these categories. But new words may lead to create new categories. An empirical approach then consists in going back and forth between a refinement of the current categorization and the assignment of corpus words to these categories.

This experiment shows that the linguistic analysis performed by ZELLIG evidences relevant classes and subclasses of lemmas. The graphs of shared contexts yield a first map of the domain. They cluster lemmas belonging to the same category. Together with the examination of graph edges, they help subcategorize some categories, and provide a valuable tool for textual exploration and analysis, akin to a synthetic concordance.

The approach works well for the most frequent words of the corpus: for instance, 19 of the 20 most frequent topography words are found in CCs. Working on a larger corpus might help categorize a greater number of lemmas. However, the required corpus size is much smaller than those (*e.g.*, 842 megabytes[3]) used by other approaches:[3,9] it is therefore less costly and easier to gather the necessary data, to analyze it and to interpret the clusters. Besides, the most frequent words are the most important ones to capture in a given specialty: 5-threshold CCs include 7 % of NP lemmas, but cover 37 % of the total surface of corpus NPs (table 1). A possible approach to categorize an even larger proportion of a corpus would be to inject identified categories into ZELLIG's process for finding second-order affinities. Basili et al.[12] show that such a technique drastically reduces frequency requirements.

Let us finally stress that the mere examination of clusters and contexts is not sufficient to determine the set of categories or their limits. On the contrary, one needs to call on domain knowledge (such knowledge can be provided by a domain expert or by an existing categorization). The incompleteness of the method is one reason. Another, more fundamental one, is that medical narratives leave implicit information that belongs to the body of knowledge shared by their authors and readers. An automated method such as that presented here therefore needs to include an interpretation step.

Acknowledgments

We thank Dr. RA Côté for graciously providing us a copy of the French version of the SNOMED Microglossary for Pathology. Serge Heiden (ELI) wrote the GraphX interactive graph handling tool, which we used to display and layout the graphs (`ftp://mycroft.ens-fcl.fr/pub/graphx/`).

References

1. Cimino JJ. Coding systems in health care. In: van Bemmel JH and McCray AT, eds, *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*. Schattauer, Stuttgart, 1996:71–85.
2. Chute CG, Cohn SP, Campbell KE, Oliver DE, and Campbell JR. The content coverage of clinical classifications. *J Am Med Informatics Assoc* 1996;3(3):224–33.
3. Hersh WR, Campbell EH, Evans DA, and Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. In: Cimino JJ, ed, Proc AMIA Annual Fall Symposium, Washington DC. AMIA, 1996:159–63. JAMIA supplement.
4. Habert B, Naulleau E, and Nazarenko A. Symbolic word clustering for medium-size corpora. In: Proc. 16th COLING, Copenhagen. 1996:490–5.
5. Harris Z, Gottfried M, Ryckman T, et al. *The Form of Information in Science, Analysis of Immunology Sublanguage*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 1989.
6. Zweigenbaum P and Consortium MENELAS . MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine* 1994;45:117–20.
7. Miller GA, Beckwith R, Fellbaum C, Gross D, and Miller KJ. Introduction to WordNet: An on-line lexical database. *Int J Lexicography* 1990;3(4).
8. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network. In: Bhargava B, Finin T, and Yesha Y, eds, Proc Second Int Conf on Information and Knowledge Management. ACM, 1993:67–74.
9. Hindle D. Noun classification from predicate-argument structures. In: Proc 28^{th} ACL, 1990:268–75.
10. Grefenstette G. Corpus-derived first, second and third-order word affinities. In: EURALEX, Amsterdam. August 1994.
11. Bensch PA and Savitch WJ. An occurrence-based model of word categorization. *Annals Math and Artif Intell* 1995;14:1–6.
12. Basili R, Pazienza MT, and Velardi P. Integrating general-purpose and corpus-based verb classification. *Comput Ling* 1996;22(4):559–68.
13. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, and Brochu L, eds. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield, 1993.
14. Bourigault D. An endogeneous corpus-based method for structural noun phrase disambiguation. In: Proc 6^{th} EACL, Utrecht. 1993:81–6.