Syntagmatic Behaviors of Verbs in Medical Texts : Expert Communication vs. Forums of Patients

Ornella Wandji Tchami, Natalia Grabar
STL UMR 8163 CNRS, Université Lille 3
59653 Villeneuve d'Ascq, France
ornwandji@yahoo.fr
natalia.grabar@univ-lille3.fr

Ulrich Heid
IWIST, Universität Hildesheim
Germany
heidul@uni-hildesheim.de

Abstract

In this paper, we propose an automatic contrastive analysis of the behavior of verbs, with regard to the semantic features of their arguments (subject, direct object, indirect object), within and across medical subcorpora. We compare four medical subcorpora with texts whose authors and intended readership have different levels of expertise. The semantic annotation of the subcorpora is based on semantic information provided by a medical terminology. Our results indicate that the proposed procedures and tools could be used for the automatic detection of different ways of expressing medical concepts and conceptual relations, according to the types of texts.

1 Introduction

Research has shown that despite the growing body of literature available to patients, communication between medical practitioners and patients is not always easy and successful. This situation is to some extent due to linguistic complexity in medical care texts (Putz (2008)). Indeed, the availability of medical information does not guarantee its readability and correct understanding. Standard medical language contains specific terminology and specialised phraseology which is hard to understand for non-expert users (McCray (2005), Zeng-Treiler et al. (2007)), and which can therefore render the communication difficult (Jucks and Bromme (2007), Tran et al. (2009)). Research into this issue has been conducted in sociology (Kharrazi (2009), Chy et al. (2012)), in Medical Informatics (Kokkinakis and Toporowska Gronostaj (2006), Smith and Wicks (2008)) and in Natural Language Processing (Zeng-Treiler and Tse (2006), Chmielik and Grabar (2011)) in order to identify the specificities of this communication. As one could expect, these studies suggested the simplification of the medical doctors' vocabulary. Researchers in NLP went further, proposing the creation of lexicons which relate expert terminology with expressions used by lay people (Zeng-Treiler and Tse (2006), Deléger and Zweigenbaum (2008), Grabar and Hamon (2014)).

In line with the studies mentioned above, we are interested in the written communication between medical experts and non-experts. We propose a comparative analysis of the distributions of argument structures (and semantic patterns) in French medical texts which have been classified and grouped according to their discursive specificity (Pearson (1998)) and the respective level of expertise of the target public. More specifically, we compare verbal arguments in four types of subcorpora, focusing on lexical preference and making different hypotheses. We assume that medical experts use more specific and specialised verbal configurations (frames, co-occurrences, collocations (i.e preferred co-occurrences)) in order to express medical concepts and the relations between them, while non-experts tend to use less specific configurations. Also we verify to which extent the semantic categories of the Snomed terminology allow to distinguish these different configurations. Our study is an extension to a previous work where we looked at the syntactic and semantic features of the elements surrounding the verbs in the expert and forum subcorpora, without taking into consideration the intermediary subcorpora and the dependency relationships between the verbs and their arguments. This work is intended to highlight the relationship between verbal argument structures and the different ways of expressing specialised concepts in texts written by people who have different levels of specialised medical knowledge. In fact, lexical preferences, collocations, semantic category preferences and verb frames share the ability to express concepts and/or relations between concepts.

2 Studies of argument structures in corpora

Investigations into the distribution of argument structures of verbs have helped describe and understand the relationship between the verbs, the argument structures they occur in and the semantic classes to which they belong. These studies have shown the tendency of particular verbs to select a particular type of arguments, and the attraction of certain argument structures for particular verbs (Gries and Stefanowitsch (2004), Gries and Stefanowitsch (2010)). Some studies focusing on verb valency patterns and their frequencies have revealed that verbs show certain preferences with respect to their valency schemes and alternations (Köhler (2005), Engelberg (2009), Cosma and Engelberg (2013)). Other researchers have automatically induced verb classes from data on the distribution of valency patterns (Schulte im Walde (2003), Schulte im Walde (2009)).

Quantitative data on argument structures are also used for the construction of lexical classes, or to build a lexical organisation which predicts much of the behaviour of a new word by associating it with an appropriate class. As far as English is concerned, several studies were conducted for the acquisition of subcategorisation information from raw corpora (Briscoe and Carroll (1997); Preiss et al. (2007)). Some of these studies like Korhonen and Briscoe (2004) use subcategorisation frames for the extension of lexical-semantic classifications. Others use them as main features for the classification of verbs in specialised texts from the biomedical domain (Korhonen et al. (2008)). Only recently, French has become the target of such research. Chesley and Salmon-Alt (2006) carried out an exploratory study of 104 common verbs that allowed them to identify 27 subcategorisation schemes. More recently, Messiant et al. (2010) have implemented a method to automatically acquire a syntactic lexicon of subcategorisation frames for French verbs from large corpora.

It has been shown that the neighborhood of a verb can be different according to the type of text in which the verb appears (Helbig (1985), Wandji Tchami et al. (2013), Wandji Tchami and Grabar (2014)). Roland and Jurafsky (1998) analyse how the frequency of verb subcatgorisation schemes is affected by corpus choice. This study has revealed that verb senses are closely related to types of discourse, in such a way that both determine the frequency of the different subcatgorisation schemes of the verbs in the corpora.

Although they all look at verbal argument structures within different types of texts, none of the above-mentioned studies proposes the kind of approach we are trying to develop. We propose a study of subcategorisation schemes in medical corpora that are differentiated according to their levels of specialization, and we use a medical terminology for the semantic annotation of the texts, to detect selectional restrictions and lexical preferences.

3 Material

The study is based on two types of material: corpora distinguished by the levels of expertise of their authors and intended readers (section 3.1) and a semantic resource (section 3.2), used for the semantic annotation of the corpora.

3.1 Corpora

The corpus is made up of a set of four medical subcorpora of written French, which are distinguished by their discursive specificities (Pearson, 1998) and the respective levels of expertise of their readership. The first three subcorpora come from the portal CISMeF[1], which indexes medical texts according to three different categories: texts for medical experts, texts for medical students, texts for patients or non-experts. The fourth subcorpus is made of texts written by non-experts. It contains discussions between patients and/or persons participating in a forum called *Doctissimo, Hypertension, Problèmes Cardiaques (Doctissimo, Hypertension, heart problems)*[2].

Corpus	Size	Verb occ.	pron. occ.	description
C_1 / *expert*	1,285,665	52529	1349	scientific publications and reports
C_2 / *student*	384,381	22092	920	didactic supports created for students
C_3 / *patient*	253,968	19421	1176	documentation and brochures
C_4 / *forum*	1,588,697	184843	8261	forum messages from participants

Table 1: Size of the subcorpora used

Table 1 indicates the size of the four subcorpora (number of tokens) and the number of verbal oc-

[1] *http://www.cismef.org/*

[2] *http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm*

currences per subcorpus; the rightmost column indicates how many verbal occurrences per subcorpus have pronominal arguments (which will not be resolved and thus not counted in this study). As can be seen, the expert and forum corpora are almost equal in size, while the student and the lay persons' corpora are much smaller, but also similar in size. We make the assumption that the authors of the four subcorpora represent actors of the medical domain, who have different levels of expertise as far as the use of specialised medical language is concerned.

3.2 Semantic resource

We use the *Snomed International Terminology* (Côté (1996)) which groups medical terms into eleven semantic categories, of which nine are considered in this study[3]. This terminology was chosen because it is one of the largest medical terminologies available for French.

$\mathcal{T}$: Topography or anatomical locations (*e.g.*, *coeur (heart), cardiaque (cardiac), digestif (digestive), vaisseau (vessel)*);

$\mathcal{S}$: Social status (*e.g.*, *mari (husband), soeur (sister), mère (mother), ancien fumeur (former smoker), donneur (donor)*);

$\mathcal{P}$: Procedures (*e.g.*, *césarienne (caesarean), transducteur ultrasons (ultrasound transducer), téléexpertise (tele-expertise)*);

$\mathcal{L}$: Living organisms, such as bacteries and viruses (*e.g.*, *Bacillus, Enterobacter, Klebsiella, Salmonella*); plants (*e.g.*, *fougère (fern), pomme de terre (potato)*), but also animals (*e.g.*, *singe (monkey), chien dalmatien (dalmatian dog)*);

$\mathcal{J}$: Professional occupations (*e.g.*, *équipe de SAMU (ambulance team), anesthésiste (anesthesiologist), assureur (insurer), magasinier (storekeeper)*);

$\mathcal{F}$: Functions and dysfunctions of the organism (*e.g.*, *pression artérielle (arterial pressure), métabolique (metabolic), protéinurie (proteinuria), détresse (distress), insuffisance (deficiency)*);

$\mathcal{D}$: Disorders and pathologies (*e.g.*, *obésité (obesity), hypertension artérielle (arterial hypertension), cancer (cancer), maladie (disease)*);

$\mathcal{C}$: Chemical products (*e.g.*, *médicament (medication), sodium, héparine (heparin), bleu de méthylène (methylene blue)*);

$\mathcal{A}$: Physical agents and artefacts (*e.g.*, *cathéter (catheter), prothèse (prosthesis), tube (tube)*).

In our approach, the semantic categories of the Snomed International terminology are considered as ontological categories used for the characterisation of the verbal arguments. The used version of Snomed contains 144 267 entries (mainly French nouns, noun phrases and adjectives). We used it for the semantic annotation of our corpus. The Snomed entries may not necessarily cover all domain notions in our texts (Chute et al., 1996). For this reason, in a previous study, we attempted to complete the coverage of the terminology in relation with the corpus used (Wandji Tchami and Grabar (2014)). We computed the plural forms of Snomed's single word terms, and we tried to detect misspellings of the terms by means of the string edit distance (Levenshtein, 1966). In both cases, the computed forms inherit the semantic type of the terms from the Snomed. In this way, 14 035 entries were added to the terminology.

4 Method

The method applied in this study aims at describing and comparing the argument structures of verbs in different types of subcorpora, with a particular focus on selectional restrictions and lexical preferences. The tools and procedures used allow us to detect collocations and different ways of expressing concepts and conceptual relations. In order to achieve our aim, we follow 3 main steps: the corpus pre-processing and annotation (syntactic and semantic) (section 4.1), the extraction of verbal argument structures and co-occurrence data (section 4.2), both performed automatically and followed by a manual analysis (section 4.3) which aims at contrasting and interpreting the automatically extracted data.

4.1 Corpus pre-processing and annotation

The subcorpora have all been downloaded from the above-mentioned online sources, converted into plain text and recoded in UTF-8 format. The syntactic analysis of sentences is performed with the Cordial dependency parser (Dominique et al., 2009). Its output contains sentences in a tabulated

[3]The two semantic classes containing modifiers are not taken into consideration in this study.

format similar to the CONLL format (Buchholz and Marsi, 2006). In this format, a sentence consists of one or more tokens, each one annotated with thirteen fields, separated by a tab character. Among these fields, the *syntactic function* and the *pivot* verb are the main information that allow us to extract the verbs and their arguments.

The syntactically annotated sentences are then processed with Perl programs that perform the semantic annotation by projecting the resource described in Section 3.2 onto the lemmatised sentences. The categories of the terminology add semantic information to the syntactic patterns of verbs. Hence, at the end of this stage, each verb argument appearing in the terminology is labeled with a semantic category, in addition to its syntactic function; such pair constitutes what we call *a specialised configuration or frame* while a pair whose argument has no Snomed categories is considered as a *non specialised configuration or frame*.

4.2 Extraction of verbal argument structures and of verb+noun co-occurrence

The sets of sentences annotated at the previous step are processed with Perl programs that extract argument structures involving the Snomed categories of terms, when provided by Snomed, as in Table 2 (*V+Su/Scat+DO/Scat, V+Su/Scat+DO/Scat+IO/Scat*) and pairs of *V+Su/Scat*, *V+DO/Scat* and *V+IO/Scat*[4].

For each verb, the most frequent cooccurring objects are automatically extracted and their corresponding frequencies are computed from all subcorpora. Indeed, in 5.1 and 5.2, we focus particularly on direct objects, except with the verb *exposer* for which we have considered the subject (*patientS+exposer*) and the indirect object (*exposer un risque*) (Table 2).

For a given verb A, after extracting its most frequent objects from the corpora, we automatically extract further verbs that frequently combine with A's objects, most particularly those which are semantically close to A, and we compute the frequency of all verb+Object pairs (see Tables 3 and 4). These data function as indicators of the phenomena observed on the medical language of experts and non experts. Indeed, this experiment helps to identify semantic groups of verbs expressing similar concepts and conceptual relations between the verb arguments.

After processing all the verbs found in the different subcorpora, 11 verbs were selected for a more detailed case study : augmenter (*add*), évaluer (*evaluate*), exposer (*expose*), subir (*undergo*), prescrire (*prescribe*), provoquer (*provoke*), accompagner (*accompany*), suivre (*follow*), causer (*cause*), baisser (*lower*), and entraîner (*lead to*). These verbs were selected according to two main criteria:

- Frequency: the verbs should have at least 20 occurrences each, in at least two of the subcorpora;
- Types of verbs: we tried to choose not only verbs that intuitively tend to have specialised usages in specialised domain texts, but also general language verbs like *accompagner*, *baisser*, *suivre* etc.; The tendency to co-occur frequently with particular terms was also taken into consideration, since we focus on lexical preference and collocation.

4.3 Comparative analysis of verbal behaviors

The comparative analysis is done manually and aims at highlighting the differences and similarities of the subcorpora with regard to selectional restrictions and lexical preferences. We compare the frequency of verbal configurations (pairs of verb+argument or frames) across the subcorpora. This analysis addresses different aspects : the arguments (terms) cooccurring with verbs, the verbs cooccurring with those arguments, the different frames verbs frequently appear in, and argument structures expressing similar conceptual relations. The results are discussed in Section 5.1.

5 Results and Discussion

5.1 Terms cooccurring with verbs

The data provided in Table 2 lead to several observations. Some verbs frequently select terms from a particular Snomed category, mostly specific terms, in a particular subcorpus, while in the other subcorpora this co-occurrence never happens or only happens scarcely. This phenomenon is particularly striking with verbs like *prescire* and *subir*. In the forum and sometimes in the lay subcorpus, these verbs frequently combine with

[4] V=verb, Su=sujet, DO, direct Object, IO=indirect Object Scat=Snomed category

Verbs	Nominal cooccurrents				
	Arguments	exp	stu	lay	for
prescrire	*traitement* $\mathcal{P}$	3	0	0	7
	examen $\mathcal{P}$	0	0	2	7
	médicament $\mathcal{C}$	0	0	7	26
subir	*ablation* $\mathcal{P}$	0	0	0	39
	intervention $\mathcal{P}$	6	0	1	30
	AVC $\mathcal{D}$	0	0	2	12
augmenter	*tension* $\mathcal{F}$	0	0	7	14
	risque/risque de $\mathcal{F}$	26	8	5	7
baisser	*tension* $\mathcal{F}$	0	0	4	18
exposer	*à+risque* $\mathcal{F}$	14	8	0	3
	patient $\mathcal{S}$	23	5	1	0
suivre	*apparition de symptômes* $\mathcal{F}$	5	0	0	0
	patient $\mathcal{S}$	6	0	0	0
	régime $\mathcal{F}$	1	0	0	5
	conseil	0	0	4	10
	traitement $\mathcal{P}$	2	2	1	13
évaluer	*patient* $\mathcal{S}$	7	0	0	0
	indication	6	0	0	0
	risque $\mathcal{F}$	9	2	0	1

Table 2: Most frequent verb/arg pairs: capital letter=the Snomed category, no capital letter=no category provided

terms belonging to category P (procedures); more specifically, *prescrire* seems to have an attraction for the terms *traitement* and *examen*, while *subir* has a strong attraction for *intervention* and *ablation* (which refers to a type of medical intervention (hyponym)). *Prescrire* also combines frequently with names of chemical products (C) and shows a particular attraction for the term *médicament*, while *subir* prefers terms referring to disorders and diseases (D), and more precisely the term *AVC* (*stroke*). These are preferred co-occurrences which are therefore seen as collocations.

Such collocations may involve polysemous verbs and their different readings. For example, in the expert subcorpus (and sometimes in the student subcorpus), *évaluer* and *suivre* tend to appear frequently with terms referring to functions of the organism (F) or to Social status (S). *Évaluer* seems to be attracted by *risque*, *indication* and *patient*. *Évaluer*+F means *to mesure, determine, calculate, gauge, quantify*, while *évaluer*+S means *to examine*.

The differences in verb/arg pair frequencies can lead to different interpretations. First of all, when the frequency difference is very important from the forum subcorpus to the expert subcorpus, this may signal some specificities of the laypersons' language. Indeed, while health care specialists share foundational domain knowledge based on formal education and professional experience, the patients' or non experts' medical language is characterised by the use of common expressions and collocations, sometimes involving technical medical terms (*prescrire un médicament*, *subir une ablation*, *subir un AVC*, *suivre un régime*) borrowed from the medical experts' language. According to researchers in Consumer Health Literature, such mixed phraseology is the result of social and cultural influence on language and they are acquired from formal and informal sources such as the internet (Zeng-Treiler et al. (2006), Zeng-Treiler and Tse (2006)). The frequent use of these expressions makes them progressively become part of everyday language. This could be a plausible explanation for the high frequency of expressions like *prescrire un médicament*, *subir une ablation* or *subir un AVC*, in the forum texts.

Secondly, looking at the results from the expert subcorpus to the forum subcorpus, we notice that sometimes the frequency difference is not very important. The explanation given above could once more apply here. Indeed, medical technical terms are quite often used by non-experts to describe medical concepts. On the other hand, when a verbal combination involving a particular Snomed category is very frequent in the expert subcorpus like *exposer* + *name of a medication* (*votre patiente est exposée au ramipril*), *évaluer* + *function* (*évaluer un risque*) while the verb is totally absent or very rare in the other subcorpora, we might deal with a highly specialised (expert) or expert language-specific usage of the verb.

5.2 Lexical preferences of the arguments for verbs

The results of Section 5.1 give an account of the lexical preferences of the verbs within and across the subcorpora. In this section, we investigate the lexical preferences of nominals in the expert and forum subcorpora. Tables 3 and 4 give the results of this experiment. These data were obtained as described in Section 4.2. The blue color represents the processed verb, the entries in the column *Arguments* are the most frequent arguments of the processed verb, and the red color represents a semantic group of verbs frequently combining with the corresponding argument in the given corpus. The numbers in bracket show the frequency of each pair verb+arg.

Depending on the corpus, certain terms frequently combine with particular verbs, in order to express a particular concept. For instance, as we can see in Table 2, the terms *médicament* and *traitement* are *prescrire's* favourite cooccurrents

Arguments	Verbal cooccurrents	
	Expert	Forum
médicament	indiquer(3), recommander(2) proposer(2)	prescrire
traitement	proposer(8), envisager(7) recommander(3), imposer(3)	
examen	imposer(1), proposer(1) recommander(1), autoriser(1)	
intervention	-	subir
ablation	faire(1)	
AVC	prsenter(4), faire(2), avoir(2)	
tension	-	baisser
régime	-	suivre
conseil	considrer(1)	
traitement	recevoir(12), bénéficier(6) faire(6), poursuivre(3),	
tension	-	augmenter

Table 3: Lexical preferences of arguments in the expert subcorpus.

Arguments	Verbal cooccurrents	
	Forum	Expert
patient	traiter(1), voir(1)	suivre
apparition de	-	
symptome	expliquer (5)	
risque	mesurer(1), juger(1), exposer(23)	évaluer
patient	-	
indication	apprécier(1)	
risque	accroître(3), multiplier(2) élever(1),	augmenter

Table 4: Lexical preferences of arguments in the forum subcorpora.

in the forum and sometimes in the lay subcorpus, while in the expert subcorpus, the terms frequently co-occur with the verbs *indiquer*, *recommander*, *proposer*, and *envisager*, *recommander*, *proposer*, *imposer*, respectively.

1) Ces médicaments ne sont plus recommandés en première intention dans le traitement de l'hypertension (*These drugs are no longer recommended as first-line in the treatment of hypertension*)

Although the two groups of verbs combine with the same terms, in the professionnal language, these verbs are not semantically equivalent, they correspond to different levels of evidence. Indeed, they are used by medical experts to express the relevance of prescribing a given drug or treatment for a given disease. In contrast, patients just know about the drug or treatment they have been prescribed for their disease but do not necessarily know about these distinctions. These examples highlight a very relevant difference in the way experts and non-experts use verbal configurations : the first choose very specific and technical configurations while the others use more general ones.

In the expert subcorpus, several sentences are in the passive voice with an omitted agent, as in Example 1. This applies to some of the above-mentioned verbs and is quite reccurrent with other verbs.

The lexical choice difference within subcorpora does not only concern terms. Verbs also select particular terms to combine with, depending on the subcorpora. For example, in the forum subcorpus, the verb *suivre* frequently co-occurs with the term *conseil*, while in the expert subcorpus, the term *conseil* does not combine with this verb. Instead, *suivre* combines with *indication*. The latter and mainly the term *recommandation*, which are semantically close to *conseil*, are very frequent in the expert subcorpus. They appear in positions where *conseil* could appear. For example, *recommandation* is combined with verbs like *proposer (4), appliquer (4), actualiser (8), publier (4), élaborer (2)* and *faire (3)*. This seems to show that the experts prefer to talk about *recommandations* and *indications* which have specific and technical meanings, while laypersons are more familiar with the term *conseil* which is a common word.

Another observation was made based on the experiment carried out. In the forum subcorpus *baisser* and *augmenter* frequently co-occur with the term *tension* (*augmenter la tension* (*increase blood pressure*), *baisser la tension* (*reduce blood pressure*) (see Table 2)), expressing different states of the blood pressure. In the expert subcorpus, none of these collocations were found. In addition, among the verbs combining with *tension* in the expert subcorpus, none is semantically related to the two verbs. However, we have noticed the presence of verb based nominalisations, constructions requiring support verbs or relational adjectives, which are synonymous with the two above-mentioned collocations : *élévation tensionnelle (4)*, and *hausse de tension (1)* correspond to *augmenter la tension*, while *réduction tensionnelle (2)*, *abaissement tensionnel (2)* and *baisse de tension (4)* have the same meaning as *baisser la tension*.

This phenomenon is consistent with the results obtained in a previous study (Wandji Tchami and Grabar (2014)) and with Condamines and Bourigault (1999)'s findings which confirmed the fact that nominal entities tend to be more frequent in expert texts than in non-expert texts. The above data demonstrate that the difference between the expert and forum texts does not lie in verbs alone, but mostly in the different types of constructions

the verbs are involved in (support verb, paraphrase, verb-based nominalisation, etc.).

5.3 Verbal frames and conceptual relations

Table 5 shows frames which represent different ways of expressing the cause-effect conceptual relation. The data were extracted from the subcorpora, through the analysis of frames of *accompagner*, *causer*, *provoquer*, and *entraîner* which are causative verbs. We are aware of the fact that some of the numbers presented in this table are not high enough to draw conclusions. However, we found it important to report them because they might highlight phenomena that could be further analysed in future work, with more data.

verbs	accompagner		causer		provoquer		entraîner	
frames	pro	for	pro	for	pro	for	pro	for
$C\_s\ D\_do$	1	0	2	1	0	11	3	1
$C\_s\ F\_do$	1	0	0	1	1	5	3	0
$D\_s\ D\_io$	5	3	0	0	0	0	0	0
$C\_s\ D\_do$	3	0	3	5	0	10	6	0
$D\_s\ F\_do$	5	1	1	1	3	8	3	0
$F\_s\ F\_do$	4	3	4	5	0	32	3	2
$F\_s\ D\_do$	1	0	1	0	0	12	5	1
$F\_s\ P\_do$	0	2	1	0	3	0	2	1
$P\_s\ F\_io$	5	0	0	0	0	0	0	0
$P\_s\ D\_do$	2	0	0	0	2	2	4	7
$P\_s\ F\_do$	0	0	0	0	1	0	5	0
$P\_s\ P\_do$	0	1	0	1	0	0	5	0
$F\_s\ F\_io$	6	3	0	0	0	1	0	0

Table 5: Frames: *s*=subject, *do*=direct object, *io*=indirect object; capital letters=Snomed semantic categories.

Many frames were identified, Table 5 shows the most frequent ones which are : $F\ D$, $P\ F$, $F\ F$, $P\ D$, $D\ F$, $F\ P$, $F\ D$, $P\ P$, $C\ D$, $D\ D$. These frames are all found in the four subcorpora but they tend to choose specific verbs depending on the subcorpus. The difference mostly lies on the lexical level with the choice of verbs. In the above-mentioned frames, the left side semantic class provokes or entails an effect or consequence that is expressed by the right side category. Let us take for example the relation *Functions-Functions* ($F\ F$), where a function of the organism has an effect on another function of the organism.

2) Exp: la prise de poids$_F$ s'accompagne d'une élévation de la pression artérielle$_F$ (*weight gain is followed by a rise in blood pressure*$_F$)

3) For: une diaphorèse$_F$ intense accompagne souvent la douleur$_F$ (*the pain*$_F$ *is often followed by an intense diaphoresis*$_F$)

4) For: le stress$_F$ provoque des spasmes vasculaires$_F$ (*stress*$_F$ *causes vascular spasms*$_F$)

As we can see from the data provided in Table 5, in the expert subcorpus, this conceptual relation is frequently expressed with the verbs *accompagner* and *entraîner* while in the forum texts the verbs *provoquer* and *causer* are the most used. This remark also applies for the other above-mentioned frames. Collocational differences between expert and forum verb use also involve differences in valency and syntactic construction. In Example 2, the verb *accompagner* is in a pronominal form with a reflexive pronoun *se/s'*; this construction is the most used one in the expert subcorpus, and in the table, it is represented by the presence of the indirect object in the frame.

Another tendency observed in the expert subcorpus is the frequent use of the passive voice with a syntactically omitted agent, while in the forum subcorpus, the active voice is the most used. This observation was already underlined in Section 5.2 with *recommander, indiquer* and *proposer*.

6 Conclusion and Perspectives

In this study, we have proposed a method for the comparative analysis of verbal argument structures in medical subcorpora whose authors and intended readership have different levels of expertise, with a focus on lexical preference. The main difference observed is that medical experts tend to choose verbal configurations with very specific and technical meanings which apply to specific situations, while non-experts use more generic and common verbal configurations. Lexical choice differences often come with differences in the syntactic constructions used. Indeed, medical expert writings are characterized by the frequent use of a passive form with an omitted agent. The analysis of the two intermediary subcorpora shows that the expert and student subcorpora are close to each other while the lay subcorpus is close to the forum. As far as the method is concerned, the use of a dependency parser seems to improve the results. However, a detailed evaluation of the parsing quality is still to be done. We are also planning to carry out the analysis exemplified here on more verbs.

References

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *In Proceedings of the ACL*, pages 356–363.

Sabine Buchholz and Erwin Marsi. 2006. Conll-xshared task on multilingual dependency parsing. In *In Proc. of CoNLL*, pages 149–164.

Jolanta Chmielik and Natalia Grabar. 2011. Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques. *TAL*, 51(2):151–179.

Christopher G. Chute, SP Cohn, KE Campbell, DE Oliver, and JR Campbell. 1996. The content coverage of clinical classifications. for the computer-based patient record institute's work group on codes & structures. *J Am Med Inform Assoc*, 3(3):224–33.

Anne Condamines and Didier Bourigault. 1999. Alternance nom/verbe : explorations en corpus spécialisés. In *Cahiers de l'Elsap*, pages 41–48, Caen, France.

Ruxandra Cosma and Stefan Engelberg, 2013. *Subjektsätze als alternative Valenzen im Deutschen und Rumänischen*.

Roger A. Côté, 1996. *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Louise Deléger and Pierre Zweigenbaum. 2008. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, pages 146–50.

Laurent Dominique, Sophie Nègre, and Patrick Séguéla. 2009. L' analyseur syntaxique Cordial dans Passage. *Actes de TALN*, 9.

Natalia Grabar and Thierry Hamon. 2014. Automatic extraction of layman names for technical medical terms. In *ICHI 2014*, Pavia, Italy.

Stefan Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis. a corpus-based perspective on "alternation". *IJCL*, 9(1):97–129.

Gerhard Helbig. 1985. Valenz und kommunikation (ein wort zur diskussion). *Deutsch als Fremdsprache*, 22:153–156.

Regina Jucks and R. Bromme. 2007. Choice of words in doctor-patient communication: an analysis of health-related internet sites. *Health Commun*, 21(3):267–77.

Hadi Kharrazi. 2009. Improving healthy behaviors in type 1 diabetic patients by interactive frameworks. In *AMIA*, pages 322–326.

Reinhard Köhler. 2005. Quantitative untersuchungen zur valenz deutscher verben. *Glotometrics*, 9:13–20.

Dimitrios Kokkinakis and M Toporowska Gronostaj. 2006. Comparing lay and professional language in cardiovascular disorders corpora. In James Cook University Pham T., editor, *WSEAS Transactions on Biology and Biomedicine*, pages 429–437.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *Proc. of COLING*, pages 449–456.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 707(10).

Alexa McCray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

Cédric Messiant, Kata Gábor, and Thierry Poibeau. 2010. Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français. *TAL*, 51(1):65–96.

Jennifer Pearson. 1998. *Terms in Context*. John Benjamins, Amsterdam/Philadelphia.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*, volume 45, page 912.

Magdalena Putz. 2008. Approaching linguistic complexity in medical care. *International Journal of Anthropology*, 23(3-4):275–284.

Douglas Roland and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of ACL*, Montreal, Quebec, Canada.

Schulte im Walde. 2003. Experiments on the automatic induction of german semantic verb classes. Technical report, Universität Stuttgart.

Catherine Smith and PJ Wicks. 2008. PatientsLikeMe: Consumer health vocabulary as a folksonomy. In *Proceedings of the AMIA 2008 Symposium*, pages 682–686.

Thi Mai Tran, H Chekroud, P Thiery, and A Julienne. 2009. Internet et soins : un tiers invisible dans la relation médecine/patient ? *Ethica Clinica*, 53:34–43.

Ornella Wandji Tchami and Natalia Grabar. 2014. Towards automatic distinction between specialized and non-specialized occurrences of verbs in medical corpora. In *Proceedings of Computerm*, pages 114–124, Dublin, Ireland, August.

Ornella Wandji Tchami, MC L'Homme, and Natalia Grabar. 2013. Discovering semantic frames for a contrastive study of verbs in medical corpora. In *TIA*, Villetaneuse.

Qing Zeng-Treiler and T Tse. 2006. Exploring and developing consumer health vocabularies. *JAMIA*, 13:24–29.

Qing Zeng-Treiler, Tony Tse, Guy Divita, Alla Keselman, Jon Crowell, and Allen C Browne. 2006. Exploring lexical forms: first-generation consumer health vocabularies. In *AMIA 2006*, pages 1155–1155.