Towards Automatic Distinction between Specialized and Non-Specialized Occurrences of Verbs in Medical Corpora

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

Abstract

The medical field gathers people of different social statuses, such as students, pharmacists, managers, biologists, nurses and mainly medical doctors and patients, who represent the main actors. Despite their different levels of expertise, these actors need to interact and understand each other but the communication is not always easy and effective. This paper describes a method for a contrastive automatic analysis of verbs in medical corpora, based on the semantic annotation of the verbs nominal co-occurents. The corpora used are specialized in cardiology and distinguished according to their levels of expertise (high and low). The semantic annotation of these corpora is performed by using an existing medical terminology. The results indicate that the same verbs occurring in the two corpora show different specialization levels, which are indicated by the words (nouns and adjectives derived from medical terms) they occur with.

1 Introduction

The medical field gathers people of different social statuses, such as medical doctors, students, pharmacists, managers, biologists, nurses, imaging experts and of course patients. These actors have different levels of expertise ranging from low (typically, the patients) up to high (*e.g.*, medical doctors, pharmacists, medical students). Despite their different levels of expertise, these actors need to interact. But their mutual understanding might not always be completely successful. This situation specifically applies to patients and medical doctors who are the two main actors within the medical field (McCray, 2005; Zeng-Treiler et al., 2007). Beyond the medical field, this situation can also apply to other domains (*e.g.*, law, economics, biology). The research question is closely linked to the readability studies (Dubay, 2004), whose purpose is to address the ease with which a document can be read and understood by people, and also the ease with which the corresponding information can be exploited by the people later. As noticed, one source of difficulty may be due to the specific and specialized notions that are used : for instance, *abdominoplasty, hymenorrhaphy, escharotomy* in medical documents, *affidavit, allegation, adjudication* in legal documents, etc. This difficulty occurs at the lexical and conceptual level. Another difficulty may come from complex syntactic structures (*e.g.*, coordinated or subordinated phrases) that can occur in such documents. Hence, this difficulty is of syntactic nature. With very simple features, reduced to the length of words and sentences, the classical readability scores address these two aspects (Flesch, 1948; Dale and Chall, 1948; Bormuth, 1966; Kincaid et al., 1975). Typically, such scores do not account for the semantics of the documents. In recent readability approaches, the semantics is being taken into account through several features, such as: medical terminologies (Kokkinakis and Toporowska Gronostaj, 2006); stylistics of documents (Grabar et al., 2007; Goeuriot et al., 2007); lexicon used (Miller et al., 2007); morphological information (Chmielik and Grabar, 2011); and combination of various features (Wang, 2006; Zeng-Treiler et al., 2007; Leroy et al., 2008; François and Fairon, 2013).

We propose to continue studying the readability level of specialized documents through the semantic features. More precisely, we propose to perform a comparative analysis of verbs observed in medical corpora written in French. These corpora are differentiated according to their levels of expertise and

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

thereby they represent the patients and the medical doctors' languages. Our study focuses on verbs and their co-occurents (nouns and adjectives deriving from medical terms), and aims to investigate on the verb semantics, according to the types of constructions and to the words with which the verb occurs in the corpora. In order to achieve this, we pay a particular attention to the syntactic and semantic features of the verbs' co-occurents in the studied texts.

Our method is based on the hypothesis according to which the meaning of a verb can be influenced or determined by its context of appearance (L'Homme, 2012) and by its arguments. Indeed, various studies on specialized languages have shown that the verb is not specialized by itself (L'Homme, 1998; Lerat, 2002). Rather, being a predicative unit that involves participants called arguments, the verb can be specialized or not, depending on its argumental structure and the nature of these arguments.

In our study, the description of verbs is similar to the one performed in Frame Semantics (FS) (Fillmore, 1982), since we provide semantic information about the verbs co-occurents. The Frame Semantics framework is increasingly used for the description of lexical units in different languages (Atkins et al., 2003; Padó and Pitel, 2007; Burchardt et al., 2009; Borin et al., 2010; Koeva, 2010) and specialized fields (Dolbey et al., 2006; Schmidt, 2009; Pimentel, 2011). Among other things, Frame Semantics provides for a full description of the semantic and syntactic properties of lexical units. FS puts forward the notion of "frames", which are defined as conceptual scenarios that underlie lexical realizations in language. A frame comprises a frame evoking lexical units (ULs) and the Frame Elements (FEs), which represent the participants to the verbal process. For instance, in FrameNet (Ruppenhofer et al., 2006), the frame CURE is described as a situation that involves some specific Frame Elements, (such as HEALER, AFFLICTION, PATIENT, TREATMENT), and includes a lexical unit such as *cure*, *alleviate, heal, incurable, treat.*[1] In our approach, an FS-like modeling should allow us to describe the semantic properties of verbs. Using this framework, we will be able to highlight the differences between the studied verbs usages through their various frames and, by doing so, uncover the linguistic differences observed in corpora of different levels of expertise. However, the FS framework will be adapted in order to fit our own objectives. Indeed, the automatic annotation of the verbs co-occurents into frames will rely on the use of a terminology (Côté, 1996) which provides a semantic category for each recorded term. These categories (*e.g.*, anatomy, disorders, procedures, chemical products) typically apply to the verb co-occurents and should be evocative of the semantics of these co-occurents and the semantic properties of verbs: we consider that the semantic categories represent the frame elements which are lexically realized by the terms, while the verbs represent the frame evoking lexical units.

In a previous study, we have looked at the behavior of four verbs (*observer* (*observe*), *détecter* (*detect*), *développer* (*develop*), and *activer* (*activate*)) in medical corpora written by medical doctors by contrast to texts written by patients (Wandji Tchami et al., 2013). The results showed that in the corpus written by doctors some verbs tend to have specific meanings, according to the type of arguments that surround them. In the current work, we try to go further by enhancing our method (improved semantic annotation, automated analysis of verbs) and by distinguishing specialized and non-specialized occurrences of verbs.

In the next sections, we present the material used (section 2), the method designed (section 3). We then introduce the results and discuss them (section 4), and conclude with future work (section 5).

2 Material

We use several kinds of material: the corpora to be processed (section 2.1), the semantic resources (section 2.2), a resource with verbal forms and lemmas (section 2.3) and a list of stopwords (section 2.4).

2.1 Corpora

We study two medical corpora dealing with the specific field of cardiology (heart disorders and treatments). These corpora are distinguished according to their levels of expertise and their discursive specificities (Pearson, 1998): *Expert* corpus contains expert documents written by medical experts for medical experts. This corpus typically contains scientific publications, and show a high level of expertise. The

[1] *https://framenet.icsi.berkeley.edu/fndrupal*

corpus is collected through the CISMeF portal[2], which indexes French language medical documents and assigns them categories according to the topic they deal with (*e.g.*, cardiology, intensive care) and to their levels of expertise (*i.e.*, for medical experts, medical students or patients). *Forum* corpus contains non-expert documents written by patients for patients. This corpus contains messages from the Doctissimo forum *Hypertension Problemes Cardiaques*[3]. It shows low level of expertise, although technical terms may also be used. The size of corpora in terms of occurrences of words is indicated in Table 1. We can see that, in number of occurrences, these two corpora are comparable as for their sizes.

Corpus	Size (occ of words)
Expert	1,285,665
Forum	1,588,697

Table 1: Size of the two corpora studied.

2.2 Semantic resources

The semantic annotation of corpora is performed using the Snomed International terminology (Côté, 1996). This resource provides terms which use is suitable for the NLP processing of documents, as these are expressions close to those used in real documents. It is structured into several semantic axes:

- $\mathcal{T}$: 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 (donnor)*);
- $\mathcal{P}$: PROCEDURES (*e.g.*, *césarienne (caesarean), transducteur à ultrasons (ultrasound transducer), télé-expertise (tele-expertise)*);
- $\mathcal{L}$: LIVING ORGANISMS, such as bacteries and viruses (*e.g.*, *Bacillus, Enterobacter, Klebsiella, Salmonella*), but also human subjects (*e.g.*, *patients (patients), traumatisés (wounded), tu (you)*);
- $\mathcal{J}$: PROFESSIONAL OCCUPATIONS (*e.g.*, *équipe de SAMU (ambulance team), anesthésiste (anesthesiologist), assureur (insurer), magasinier (storekeeper)*);
- $\mathcal{F}$: FUNCTIONS 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 (*e.g.*, *prothèses (prosthesis), tube (tube), accident (accident), cathéter (catheter)*).

Further to our previous work (Wandji Tchami et al., 2013), we have added another semantic axis $\mathcal{E}$ STUDIES, that groups terms related to the scientific work and experiments (*e.g.*, *méthode* (*method*), *hypothèse* (*hypothesis*)...). Such notions are quite frequent in the corpora, while they are missing in the terminology used. The only semantic category of Snomed that we ignore in this analysis contains modifiers (*e.g.*, *aigu (acute), droit (right), antérieur (anterior)*), which are meaningful only in combination with other terms. Besides, such descriptors can occur within medical and non-medical contexts.

As stated above, we expect these semantic categories to be indicative of frame elements (FEs), while the individual terms should correspond to lexical realizations of those FEs, as in Framenet. For instance,

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

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

the Snomed category DISORDERS should allow us to discover and group under a single label terms that denote the same notion (*e.g.*, *hypertension* (*hypertension*), *obésité* (*obesity*)) related to the FE DISORDER.

The existing terminologies may not provide the entire coverage of the domain notions (Chute et al., 1996; Humphreys et al., 1997; Hole and Srinivasan, 2000; Penz et al., 2004). For this reason, we attempted to complete the coverage of the Snomed International terminology in relation with the corpora used. We addressed this question in two ways:

- We computed the plural forms for simple terms that contain one word only. The motivation for this processing is that the terminologies often record terms in singular forms, while the documents may contain singular and plural forms of these terms.

- We tried to detect the misspellings of the terms using the string edit distance (Levenshtein, 1966). This measure considers three operations: deletion, addition and substitution of characters. Each operations cost is set to 1. For instance, the Levenshtein distance between *ambolie* and *embolie* is 1, that corresponds to the substitution of *a* by *e*. The minimal length of the processed words should not be lesser than six characters, because with shorter words the propositions contain too much of errors. The motivation for this kind of processing is that it is possible and frequent to find misspelled words in real documents, especially in the forum discussions (Balahur, 2013).

In both cases, the computed forms inherit the semantic type of the terms from the terminology. For instance, *ambolie* inherits the $\mathcal{D}$ DISORDER semantic type of *embolie*. Besides, we also added the medication names from the Thériaque resource[4]. These are assigned to the $\mathcal{C}$ CHEMICAL PRODUCTS semantic type. The whole resource contains 158,298 entries.

2.3 Resource with verbal forms

We have built a resource with inflected forms of verbs: 177,468 forms for 1,964 verbs. The resource is built from the information available online[5]. The resource contains simple (*consulte, consultes, consultons* (*consult*)) and complex (*ai consulté, avons consulté* (*have consulted*)) verbal forms. This resource is required for the lemmatization of verbs (section 3.3).

2.4 List of stopwords

The list of stopwords contains grammatical units, such as prepositions, determinants, pronouns and conjunctions. It provides 263 entries.

3 Method

We first perform the description of verbs in a way similar to FS and then compare the observations made in the two corpora processed. The proposed method comprises three steps: corpora pre-processing (section 3.1), semantic annotation (section 3.2), and contrastive analysis of verbs (section 3.3). The method relies on some existing tools and on specifically designed Perl scripts.

3.1 Corpora pre-processing

The corpora are collected online from the websites indicated above and properly formatted. The corpora are then analyzed syntactically using the Bonsai parser (Candito et al., 2010). Its output contains sentences segmented into syntactic chunks (*e.g.*, NP, PP, VP) in which words are assigned parts of speech, as shown in the example that follows:

> *Le traitement repose sur les dérivés thiazidiques, plus accessibles, disponibles sous forme de médicaments génériques.*
> (*The treatment is based on thiazidic derivates, more easily accessible, and available as generic drugs.*)
> *((SENT (NP (DET Le) (NC traitement)) (VN (V repose)) (PP (P sur) (NP (DET les) (NC*

[4] *http://www.theriaque.org/*

[5] *http://leconjugueur.lefigaro.fr/frlistedeverbe.php*

> *dérivés) (AP (ADJ thiazidiques) (COORD (PONCT ,) (NP (DET les) (ADV plus) (ADJ accessibles)) (PONCT ,) (AP (ADJ disponibles)))) (PP (P sous_forme_de) (NP (NC médicaments) (AP (ADJ génériques))))))))*

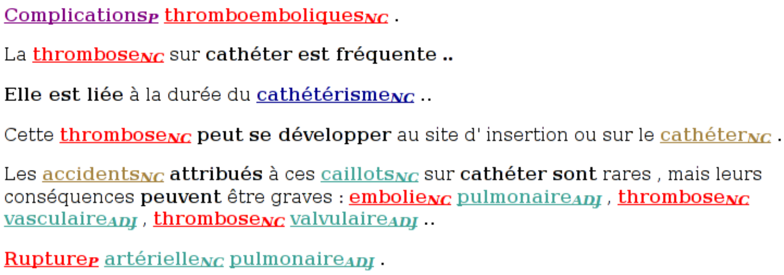
The syntactic parsing was performed in order to identify the syntactic chunks, nominal and verbal, to prepare the recognition and annotation of the terms they contain and to better the recognition of verbs. The Bonsai parser was chosen: it is adapted for french texts and it provides several hierarchical syntactic levels within the sentences and phrases. For instance, the phrase *médicaments génériques* (*generic drugs*) is syntactically analyzed as NP: *(NP (NC médicaments) (AP (ADJ génériques))))* that contains one NP *médicaments* and two APs *génériques* and the final dot. The VP of the sentence contains the verb *repose* (*is based*). As we can observe, the output of the Bonsai parser neither provides the lemmas of the forms nor the syntactic dependencies between the constituents. So our study concentrates on the verbs co-occurences with nouns, noun phrases and some relationnal adjectives. The further analysis of the corpora is based on this output.

3.2 Semantic annotation

The Bonsai format is first converted into the XML format: we work on the XML-tree structure. The semantic annotation of the corpora is done automatically. For this task, the Snomed International terminology was chosen because it is suitable for french and it offers a better outreach of the french medical language. We perform the projection of terms from the terminology on the syntactically parsed texts :

- All the chunks (NPs, PPs, APs and VPs) are processed from the largest to the smallest chunks, within which we try to recognize the terminology entries which co-occur with the verbs in the corpora. Indeed, at this stage, since our chunker does not provide dependency relations, we can only work on nouns and noun phrases that co-occur with the verbs. For instance, the largest chunk *(NP (NC médicaments) (AP (ADJ génériques))))* gives *médicaments génériques*, (*generic drugs*) that is not known in the terminology. We then test *médicaments* (*drugs*) and *génériques* (*generic*), of which *médicaments* (*drugs*) is found in the terminology and tagged with the $\mathcal{C}$ CHEMICAL PRODUCTS semantic type.
- Those VPs in which no terms have been identified are considered to be verbal forms or verbs.

Examples of corpora enriched with the semantic information are shown in Figures 1 (expert corpus) and 2 (forum corpus). In these Figures, verbs are in bold characters, semantic labels for the verbs co-occurents are represented by different colors: DISORDERS in red, FUNCTIONS in purple, ANATOMY in clear blue. These semantic categories, provided by the terminological resource, label the words that are likely to correspond to FEs.



Complications$_{P}$ thromboemboliques$_{NC}$.

La thrombose$_{NC}$ sur **cathéter est fréquente ..**

Elle est liée à la durée du cathétérisme$_{NC}$ ..

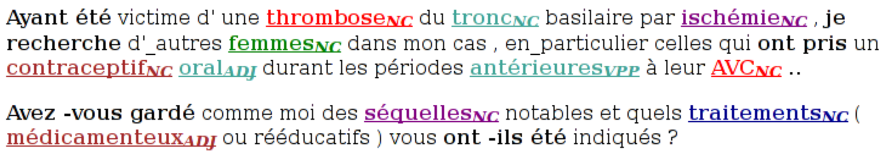
Cette thrombose$_{NC}$ **peut se développer** au site d' insertion ou sur le cathéter$_{NC}$.

Les accidents$_{NC}$ **attribués** à ces caillots$_{NC}$ sur **cathéter sont** rares , mais leurs conséquences **peuvent** être graves : embolie$_{NC}$ pulmonaire$_{ADJ}$, thrombose$_{NC}$ vasculaire$_{ADJ}$, thrombose$_{NC}$ valvulaire$_{ADJ}$ ..

Rupture$_{P}$ artérielle$_{NC}$ pulmonaire$_{ADJ}$.

Figure 1: Examples of annotations in expert corpus

We can see that in the two corpora, there are both short and long sentences. Besides, the terms recognized are often atomic. For instance, we do not recognize complex terms *embolie pulmonaire* and *thrombose du tronc*, but their simple atomic components *embolie, pulmonaire, thrombose* and *tronc*. Also, some terms match none of the terminology's entries because they are part of VPs, such as *cathéter* in Figure 1.



Ayant été victime d' une thrombose$_{NC}$ du tronc$_{NC}$ basilaire par ischémie$_{NC}$, je recherche d'_autres femmes$_{NC}$ dans mon cas , en_particulier celles qui ont pris un contraceptif$_{NC}$ oral$_{ADJ}$ durant les périodes antérieures$_{VPP}$ à leur AVC$_{NC}$ ..

Avez -vous gardé comme moi des séquelles$_{NC}$ notables et quels traitements$_{NC}$ (médicamenteux$_{ADJ}$ ou rééducatifs) vous ont -ils été indiqués ?

Figure 2: Examples of annotations in forum corpus.

3.3 Automatic analysis of verbs

For the analysis of the verbs, we extract information related to verbs and to the words with which they occur. Currently, only sentences with one VP are processed 8 842 sentences for the expert corpus and 10 563 for the forum corpus.

- *Lemmatization of verbs.* As we noticed, the syntactic parser's output does not provide the lemmas. For the lemmatization of the verbs, we use the verbal resource described in section 2.3. Hence, the content of the verbal chunk is analyzed:
  - it may contain a simple or complex verbal form that exists in the resource, in which case we record the corresponding lemma;
  - if the whole chunk doesnot appear in the resource, we check out its atomic components: if all or some of these components are known, we record the corresponding lemmas. This case may apply to passive structures (*a été conseillé* (*has been advised*)), insertions (*est souvent conseillé is often advised*) or negations (*n'est pas conseillé* (*is not advised*)): in these cases, the lemmas are *avoir être conseiller*, *être conseiller* and *être conseiller*. These lemmas will be normalized in the further step: the head verb will be chosen automatically and considered as the main lemma within the verbal phrase;
  - finally, the VPs may consist of words that are not known in the verb resource. These may be morphologically contructed verbs (*réévaluer* (*reevaluate*)) or, words from other parts of speech, errouneously considered as verbs (*e.g.*, *télédéclaration, artérielle, stroke*). This is unfortunately a very frequent case.

- *Extraction of information related to the verb co-occurents.* For the extraction of these information, we consider all the verbs appearing in sentences with one VP. For each verb, we distinguish between:
  - semantically annotated co-occurents, that are considered to be specialized;
  - and the remaining content of the sentence (except the words that are part of the stoplist), more precisely noun phrases, is considered to contain non specialized co-occurents.

  In both cases, for each verb, we compute the number and the percentage of words in each of the above mentionned categories of co-occurents.

Finally, we provide a general analysis of the corpora. For each verb, we compute: the number of occurrences in each corpus, the total, minimal, maximal and average numbers of co-occurents, both specialized and non-specialized. On the basis of this information, we analyse the differences and similarities which may exist between the use of verbs in the two corpora studied. The purpose is to provide information about the specialized and non-specialized occurrences of verbs.

4 Results and Discussion

4.1 Corpora pre-processing

The parsing, done with the Bonsai parser, provided the syntactic annotation of corpora into syntactic constituents. We have noticed some limitations:

- The Bonsai parser does not perform the lemmatization of lexical units whereas we needed to extract the verbs lemmas. The use of external resources made it possible to overcome this limitation;

- The verbal chunks do not always contain verbal constituents, but can contain other parts of speech (*e.g.*, *télédéclaration, artérielle, stroke*) and even punctuation. This is an important limitation for our work, mainly because we focus on verbs. Therefore, if we cannot extract the verbs properly, this can obviously have a negative impact on the final results. These limitations, resulting from the Bonsai parser, highlight some of the issues that characterize the state of arts as far as the syntatic analysis for French is concerned. For the future work, we are planning to try other syntactic parsers for French.

4.2 Semantic annotation

Concerning the semantic annotation we have made several observations:

- Some annotations are missing, such as *site d'insertion* (*insertion site*) that can be labeled as TOPOGRAPHY or *risque* (*risk*) as FUNCTION. This limitation is also related to the annotation of the forum corpus, that often contains misspellings or non-specialized equivalents of the terms. This limitation must be addressed in future work in order to detect new terms or the variations of the existing terms to make the annotation more exhaustive;

- Other annotations are erroneous, such as *or* (*ou*) in French annotated as CHEMICALS (*gold*)) in English-language sentences. In future, the sentences in English will be forehand filtered out at the processing stage;

- The terminological variation and the syntactic parsing provided by Bonsai make the recognition of several complex terms difficult. As we noticed previously, we mainly recognize simple atomic terms. For the current purpose, this is not a real limitation: the main objective is to detect the specialized and non-specialized words that co-occur with the verbs. Still, the number and semantic types of these words co-occuring with verbs can become biased. For instance, instead of one DISORDER term *embolie pulmonaire* (*air embolism*), we obtain one DISORDER term *embolie* (*embolism*) and one ANATOMY term *pulmonaire* (*air*).

4.3 Automatic analysis of verbs

The contrastive analysis of the words, co-occuring with verbs, provides the main results of the proposed study.

Corpus	$Total_V$	$Total_{coocc}$	$Total_{sp-coocc}$	$Total_{\neg sp-coocc}$	$A_{sp-coocc}/V$	$A_{\neg sp-coocc}/V$
Expert Ex	545	17632	8354	9272	15	17
Forum Fo	592	10852	5545	5307	9	8

Table 2: General information related to the verbs and their co-occurent words: total and average numbers of co-occurents

In Table 2, we compute the total number of verbs ($Total_V$), the total number of words co-occuring with verbs per corpus ($Total_{coocc}$), the total number of non specialized co-occurents per corpus ($N_{sp-coocc}$), the average number of specialized co-occurents per verb ($A_{sp-coocc}/V$), the average number of non specialized per verb ($A_{\neg sp-coocc}/V$). We can notice that the forum corpus provides slightly more verbs than the expert corpus. This observation might be considered to be obvious, since the forum corpus is a bit larger than the expert corpus. But if we combine this with the fact that the numbers and average numbers of co-occurents (specialized and non-specialized) are higher in the expert corpus, then the observation start making sense, since these results can be related to the confirmation by (Condamines and Bourigault, 1999) of the fact that nominal forms tend to be more frequent in specialized texts, whereas verbal forms tend to be more frequent in non-specialized texts. However, it is important to notice that some candidates in the list of non-specialized co-occurents have to be filtered out, such as adverbs (*conformément, régulièrement, précocément, partiellement*) and non relationnal adjectives (*variables, inconscients, différents*). The abundance of adverbs in the expert corpus (Table 4) by contrast to the forum corpus, where their presence seems to be less important, is consistent with the previous work, which show that non-specialized documents tend to have simpler syntactic and semantic structures (Wandji Tchami et al., 2013) and less adverbs (Brouwers et al., 2012).

Verbs	N_{occc}		N_{coocc}		$N_{sp-coocc}$		$\%_{sp-coocc}$		$N_{\neg sp-coocc}$		$\%_{\neg sp-coocc}$		$A_{sp-coocc}$		$A_{\neg sp-coocc}$	
	Ex	Fo	Ex	Fo	Ex	Fo	Ex	Fo	Ex	Fo	Ex	Fo	Ex	Fo	Ex	Fo
augmenter	21	14	122	52	62	26	51.5	56.2	60	26	48.4	43.7	2.9	1.8	2.8	1.8
causer	5	7	26	27	17	19	72	68.2	9	8	28	31.72	3.4	2.7	1.8	1.1
favoriser	10	6	56	22	38	17	70.5	77.3	18	5	29.4	22.6	3.8	2.8	1.8	0.8
prescrire	6	29	30	108	16	71	58.9	69.7	14	37	41	30.2	2.6	2.4	2.3	1.2
provoquer	7	15	60	64	32	37	57	70.2	28	27	42.9	29.7	4.5	2.4	4	1.8
risquer	7	7	18	13	12	11	1.7	1.5	6	2	0.8	0.2	78.5	90	21.42	10
signaler	12	4	73	14	32	7	46.9	48.3	41	7	53	51.6	2.6	1.7	3.4	1.7
subir	4	24	20	98	15	54	76.1	63	5	44	23.8	36.9	3.7	2.5	1.2	1.8
traiter	24	17	107	67	66	34	65	60.2	41	33	34.9	39.7	2.7	2	1.7	1.9

Table 3: Information on some verbs that occur in Expert Ex and Forum Fo corpora

In Table 3, we give similar information but for with individual verbs. For each verb, in every corpus, we compute the number of occurence (N_{occ}), the number of words (N_{coocc}) occuring with the verb, the number of specialized co-occurents ($N_{sp-coocc}$), the percentage of specialized co-occurents ($\%_{sp-coocc}$), the number of non specialized co-occurents ($N_{\neg sp-coocc}$), the percentage of non specialized co-occurents ($\%_{\neg sp-coocc}$), the average number of specialized co-occurents ($A_{sp-coocc}$) and the average number of non specialized co-occurents ($A_{\neg sp-coocc}$). These verbs are chosen because they occur in the two corpora studied and because they are sufficiently frequent as compared to others. In our opinion, these verbs may receive specialized and non-specialized meanings according to their usage. Indeed, Table 3 shows that these verbs behave differently according to the corpus. On the one hand, there are verbs (*e.g.*, *augmenter, favoriser, signaler, traiter, risquer*) that occur with an important number of specialized co-occurents in the Experts Ex corpus while they have lower numbers of specialized co-occurents in the Forum Fo corpus. On the other hand, there are verbs (*e.g.*, *causer, subir, prescrire*) that have more specialized co-occurents in the Forum corpus than in the Expert corpus. If we consider the number of occurrences of these verbs, we can definitely notice that some of them (*e.g. causer* and *subir*) regularly occur with more specialized co-occurents in the Expert corpus (although with lower number of specialized co-occurents) than in the Forum corpus. This means that their frames involve different numbers of specialized co-occurents, that are higher in the Expert corpus.

In table 4, we show the frequent co-occurents for five verbs. We can propose two main observations:

- Some verbs involve an important number of specialized co-occurents, that have different semantic types in the Expert and Forum corpora. For instance, the verb *augmenter* provides a total of 88 specialized co-occurents that belong to nine semantic types ($\mathcal{D}$, $\mathcal{P}$, $\mathcal{S}$, $\mathcal{J}$, $\mathcal{C}$, $\mathcal{F}$, $\mathcal{T}$, $\mathcal{L}$ and $\mathcal{A}$). The most frequent among them are $\mathcal{F}$ (27), $\mathcal{D}$ (18), $\mathcal{T}$ (15), and $\mathcal{P}$ (9), and occur mostly in the Expert corpus. These might be more general verbs, with weaker specific selectional restrictions.

- Other verbs frequently occur with specialized terms that belong to a specific semantic type. This most frequent label can be specific to one corpus only or simultaneously to the two. For instance, for the verb *prescrire*, the most frequent labels are the same in the two corpora: $\mathcal{C}$, $\mathcal{J}$, $\mathcal{P}$ and $\mathcal{T}$ terms. *Traiter* frequently occurs, in the two corpora, with $\mathcal{C}$ and $\mathcal{D}$ terms.

The general observation is that, for a given verb, the Expert corpus shows more sophisticated syntactic structures with higher number of specialized co-occurents. Besides, some verbs may show similar or different behavior in the two corpora studied. According to the objectives of the proposed work, we consider that an important presence of specialized terms in a sentence or corpus indicates a very specialized use and meaning of the verbs. Quantitative and qualitative analysis of the data support this first study and results.

	$sp - coocc$		$\neg sp - coocc$	
verbs	*Expert*	*Forum*	*Expert*	*Forum*
augmenter	thrombolyses/$\mathcal{F}$, gliomes/$\mathcal{D}$, O2/$\mathcal{C}$, rétinopathie/$\mathcal{D}$, Glasgow/$\mathcal{P}$, myocardique/$\mathcal{T}$	BNP/$\mathcal{P}$, infarctus/$\mathcal{D}$, lasilix/$\mathcal{C}$, mouvements/$\mathcal{A}$, tabac/$\mathcal{L}$	inférieur, égal, score, groupe, inconscients, précocément	heures, légèrement
prescrire	protocole/$\mathcal{P}$, anticoagulant/$\mathcal{C}$, BNP/$\mathcal{P}$	comprimé/$\mathcal{C}$, diurétique/$\mathcal{C}$, médecin/$\mathcal{J}$	ministre, publication, régulièrement	jour, matin, variables
produire	pression/$\mathcal{F}$, contraction/$\mathcal{D}$	spasmes/$\mathcal{F}$, coronnaires/$\mathcal{T}$, stenosees/$\mathcal{D}$	gauche, grande, onde, antérograde, différents	général, déja
traiter	hypoglycémies/$\mathcal{D}$, prévention/$\mathcal{P}$	insuffisance/$\mathcal{F}$, cardiaque/$\mathcal{T}$, anévrismes/$\mathcal{D}$	réccurentes, cas, partiellement	succès, près de, suite
provoquer	fibrose/$\mathcal{D}$, tissus/$\mathcal{A}$, nerveux/$\mathcal{T}$, Vibrio/$\mathcal{L}$, vomissements/$\mathcal{F}$	extrasystoles/T, AVC/$\mathcal{D}$, père/$\mathcal{S}$, malaise/$\mathcal{F}$, mouvement/$\mathcal{A}$	secondaires, volontairement, insatisfaisantes, relativement, peu, alimentaire, striés	différent, beaucoup, gênant, angoissant, mini, gros, longue, petite, soirée
subir	patient/$\mathcal{J}$, arthroplastie/$\mathcal{P}$	pose/$\mathcal{P}$, fibrillation/$\mathcal{F}$, AVC/$\mathcal{D}$	raison,fixateur, externe	fuite, grade

Table 4: Description of the verbs co-occurents

5 Conclusion

We have proposed an automatic method to distinguish between specialized and non-specialized occurrences of verbs in medical corpora. This work is intended to enhance the previous study (Wandji-2013). Indeed, the method used has changed from semi-automatic to completely automatic; and a new task is performed in order to enhance the annotation process : the syntactic parsing of the corpora. Also, some new materials are used namely the Bonsai parser, the resource of verbal forms, the stoplist. There is an increase in the quantity of data analyzed; all the verbs of the various corpora were considered in this study. The annotation is based on an approach similar to Frame Semantics, considering the fact that semantic information related to the verbs co-occurents are provided through the use of a medical terminology. Though our method is still under development, it has helped to notice that some verbs regularly co-occur with specialized terms in a given context or corpus while in another, the same verbs mostly occurs with general language words. This observation takes us back to the issue of text readability, described in the introduction. Indeed, the verbs whose occurences are characterized by the predominance of specialized terms, can be considered as sources of reading difficulties for non experts in medecine.

6 Future work

We plan to extend this study in different ways. The recognition of the verb neighbors must be improved with the main objective to make the annotations more exhaustive. In this study, we have portrayed the verbs behaviors and their relations with the words with which they occur in the corpora. However, our aim is to automatically identify the verbs arguments, among his co-occurents. We also plan to peform an automatic distinction between : the syntactic functions (subject, object, etc.) of the verbs arguments and the core and non-core elements. We also plan to compute the dependency relations within sentences, either by using another chunker or by integrating to our treatment chain a tool that can perform this task. In addition, we will concentrate on the description of semantic frames of the medical verbs and on the identification of other eventual reading difficulties that might be related to the verbs usages in the corpora. As indicated above, we processed sentences that have only one verbal phrase (8 842 for the Forum corpus and 10 563 for the Expert corpus). In the future, we will process other sentences, coordinated or subordinated, which will be segmented into simple propositions before the processing. Another point is related to the exploitation of these findings for the simplification of medical documents at two levels: syntactic and lexical. Finally, working at a fine-grained verbal semantics, we can distinguish the uses of verbs according to whether their semantics and frames remain close or indicate different meanings.

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) and the DGA, under the Tecsan grant ANR-11-TECS-012.

References

S Atkins, M Rundell, and H Sato. 2003. The contribution of framenet to practical lexicography. *International Journal of Lexicography*, 16(3):333–357.

A Balahur. 2013. Sentiment analysis in social media texts. In *Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128.

L Borin, D Dannélls, M Forsberg, M Toporowska Gronostaj, and D Kokkinakis. 2010. The past meets the present in the swedish framenet++. In *14th EURALEX International Congress*, pages 269–281.

J Bormuth. 1966. Readability: A new approach. *Reading research quarterly*, 1(3):79–132.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2012. Simplification syntaxique de phrases pour le français. In *TALN*, pages 211–224.

A Burchardt, K Erk, A Frank, A Kowalski, S Padó, and M Pinkal, 2009. *Using FrameNet for the semantic analysis of German: Annotation, representation, and automation*, pages 209–244.

M Candito, J Nivre, P Denis, and E Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *International Conference on Computational Linguistics*, pages 108–116.

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

CG 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.

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

E Dale and JS Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27:11–20.

AM Dolbey, M Ellsworth, and J Scheffczyk. 2006. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *KR-MED*. 87-94.

William H. Dubay. 2004. The principles of readability. *Impact Information*. Available at *http://almacenplantillasweb.es/wp-content/uploads/2009/11/The-Principles-of-Readability.pdf.*

C Fillmore, 1982. *Frame Semantics*, pages 111–137.

R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 23:221–233.

T François and C Fairon. 2013. Les apports du TAL à la lisibilité du français langue étrangère. *TAL*, 54(1):171–202.

L Goeuriot, N Grabar, and B Daille. 2007. Caractérisation des discours scientifique et vulgarisé en français, japonais et russe. In *TALN*, pages 93–102.

N Grabar, S Krivine, and MC Jaulent. 2007. Classification of health webpages as expert and non expert with a reduced set of cross-language features. In *AMIA*, pages 284–288.

WT Hole and S Srinivasan. 2000. Discovering missed synonymy in a large concept-oriented metathesaurus. In *AMIA 2000*, pages 354–8.

BL Humphreys, AT McCray, and ML Cheh. 1997. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *J Am Med Inform Assoc*, 4(6):484–500.

JP Kincaid, RP Jr Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

S Koeva. 2010. Lexicon and grammar in bulgarian framenet. In *LREC'10*.

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

P Lerat. 2002. Qu'est-ce que le verbe spécialisé? le cas du droit. *Cahiers de Lexicologie*, 80:201–211.

G Leroy, S Helmreich, J Cowie, T Miller, and W Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *AMIA 2008*, pages 394–8.

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

MC L'Homme. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2):61–84.

Marie-Claude L'Homme. 2012. Le verbe terminologique: un portrait des travaux récents. In *CMLF 2012*, pages 93–107.

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

T Miller, G Leroy, S Chatterjee, J Fan, and B Thoms. 2007. A classifier to evaluate language specificity of medical documents. In *HICSS*, pages 134–140.

S Padó and G Pitel. 2007. Annotation précise du francais en sémantique de rôles par projection cross-linguistique. In *TALN 2007*.

J Pearson. 1998. *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam/Philadelphia.

JF Penz, SH Brown, JS Carter, PL Elkin, VN Nguyen, SA Sims, and MJ Lincoln. 2004. Evaluation of snomed coverage of veterans health administration terms. In *Medinfo*, pages 540–4.

J Pimentel. 2011. Description de verbes juridiques au moyen de la sémantique des cadres. In *TOTH*.

J Ruppenhofer, M Ellsworth, MRL Petruck, C R. Johnson, and J Scheffczyk. 2006. Framenet ii: Extended theory and practice. Technical report, FrameNet. Available online http://framenet.icsi.berkeley.edu.

T Schmidt, 2009. *The Kicktionary – A Multilingual Lexical Resource of Football Language*, pages 101–134.

O Wandji Tchami, MC L'Homme, and N Grabar. 2013. Discovering semantic frames for a contrastive study of verbs in medical corpora. In *Terminologie et intelligence artificielle (TIA)*, Villetaneuse.

Y Wang. 2006. Automatic recognition of text difficulty from consumers health information. In IEEE, editor, *Computer-Based Medical Systems*, pages 131–136.

Q Zeng-Treiler, H Kim, S Goryachev, A Keselman, L Slaugther, and CA Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, pages 1117–1121, Brisbane, Australia.