Discovering Semantic Frames for a Contrastive Study of Verbs in Medical Corpora

Ornella Wandji
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq, France
ornwandji@yahoo.fr

Marie-Claude L'Homme
OLST, Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal H3C 3J7
Québec, Canada
mc.lhomme@umontreal.ca

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

Abstract

The field of medicine gathers actors with different levels of expertise. These actors must interact, although their mutual understanding is not always completely successful. We propose to study corpora (with high and low levels of expertise) in order to observe their specificities. More specifically, we perform a contrastive analysis of verbs, and of the syntactic and semantic features of their participants, based on the Frame Semantics framework and the methodology implemented in FrameNet. In order to acheive this, we use an existing medical terminology to automatically annotate the semantics classes of participants of verbs, which we assume are indicative of semantics roles. Our results indicate that verbs show similar or very close semantics in some contexts, while in other contexts they behave differently.

1 Introduction

The field of medicine is heterogeneous because it gathers actors with various backgrounds, 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). Moreover, actors with different levels of expertise interact, but their mutual understanding might not always be completely successful. This specifically applies to patients and medical doctors (AMA, 1999; McCray, 2005; Zeng-Treiler et al., 2007), but we assume that similar situations apply to other actors.

In this study, we propose to perform a comparative analysis of written medical corpora, which are differenciated according to their levels of expertise. More specifically, we concentrate on the study of selected verbs used in these corpora and aim to characterize the syntactic and semantic features of their participants. Most of the participants are arguments (or, in terms of Frame Semantics, *core frame elements*). They often correspond to noun phrases. The description of verbs is based on the Frame Semantics framework (Fillmore, 1982). We assume that verbs are an excellent starting point for modeling the contents and semantics of sentences. The study is perfomed with French data. In the following, we briefly present previous work on verbs in specialized languages (section 2) and on Frame Semantics (section 3). We also describe the material that we use (section 4) and the method developed to process it (section 5). We then give an account of the results (section 6), and conclude with some directions for future work (section 7).

2 Verbs in specialized languages

Traditionally, the study of specialized languages focuses on nominal entities (typically, nouns and noun phrases), commonly used for the compilation of terminologies, ontologies, thesauri or vocabularies. This situation can be explained by the needs raised by specific applications (*i.e.*, indexing or information retrieval are typically based on nominal entities), but it can also be explained by theoretical and methodological approaches that were designed for processing nominal entities. Nevertheless, an increasing number of researchers now address the study of verbs and of their role

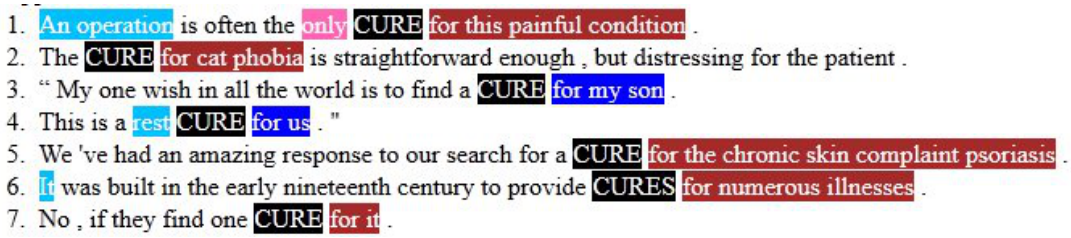

Figure 1: Example of the FrameNet annotations of the lexical unit CURE.

in specialized fields. Specific methods were developed in order to exploit verbs in terminological descriptions: in banking (Condamines, 1993), computer science (L'Homme, 1998), environment (L'Homme, 2012) and law (Lerat, 2002; Pimentel, 2011). The approaches taken by these authors differ, but they all agree on the importance of supplying a characterization of the arguments of specialized verbs. Notice also that TermoStat[1] (Drouin, 2003) can extract verbs from specialized corpora. Indeed, it has been demonstrated that verbs play an important role in Natural Language Processing (NLP) tasks, such as the detection of interactions between proteins or more generally in the extraction of semantic relations (Godbert et al., 2007; Rupp et al., 2010; Thompson et al., 2011; Miwa et al., 2012; Roberts et al., 2008).

3 Frame Semantics

The study of verbs we propose is based on Frame Semantics (FS) (Fillmore, 1982). This framework is increasingly used for the description of lexical units in different languages, mainly in English (Gildea and Jurafsky, 2002; Atkins et al., 2003; Basili et al., 2008), but it was soon extended to other languages (Padó and Pitel, 2007; Burchardt et al., 2009; Ohara, 2009; Borin et al., 2010; Koeva, 2010). Until recently, French has been neglected with regard to this framework. In addition to the description of general language, this framework can be adapted to take into account data from specialized languages (Dolbey et al., 2006; Schmidt, 2009; Pimentel, 2011). Other resources include a fine-grained characterization of the semantics and syntax of lexical units. For instance, while focussing on verbs (as opposed to FrameNet that takes into account all "frame-bearing units"), VerbNet (Palmer, 2009) implements a description of verbs and their argument structure within a similar framework.

FS puts forward the notion of "frames", which are defined as conceptual scenarios that underlie lexical realizations in language. For instance, in FrameNet (Ruppenhofer et al., 2006), the lexical database that implements the principles of FS, the frame CURE is described as a situation that comprises specific Frame Elements (FEs), (such as HEALER, AFFLICTION, PATIENT, TREATMENT, MEDICATION), and includes lexical units (LUs) such as *cure* (noun and verb), *alleviate, heal, healer, incurable, nurse, treat*.[2] In addition to the description of the frame, FrameNet provides annotations for LUs that evoke it (Figure 1).

According to our hypothesis, an FS-like modeling should allow us to describe the syntactic and semantic properties of specialized verbs and, by doing so, uncover linguistic differences observed in corpora of different levels of expertise.

4 Material

We use two kinds of material: corpora distinguished by their levels of expertise (section 4.1) and semantic resources (section 4.2), that are used for the semantic annotation of corpora.

4.1 Corpora building and processing

We study four medical corpora dealing with the specific field of cardiology. These corpora are distinguished according to their discoursive specificities and levels of expertise (Pearson, 1998). The first three corpora are collected through the CISMeF portal[3], 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), the forth corpus is extracted from the

[1] *http://olst.ling.umontreal.ca/~drouinp/termostat_web/*

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

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

Corpus	Size (occ of words)
C_1 / *expert*	1,285,665
C_2 / *student*	384,381
C_3 / *patient*	253,968
C_4 / *forum*	1,588,697

Table 1: Size of the corpora.

Doctissimo forum *Hypertension Problemes Cardiaques*[4]. The size of corpora in terms of occurrences of words is indicated in Table 1.

- C_1 or *expert* corpus contains expert documents written by medical experts for medical experts. These documents usually correspond to scientific publications and reports. They show a high level of expertise;

- C_2 or *student* corpus contains expert documents written by medical experts for medical students. These documents usually correspond to didactic support created for medical students. This corpus shows a middle level of expertise: it contains technical terms that are usually introduced and defined;

- C_3 or *patient* corpus contains non-expert documents usually written by medical experts or medical associations for patients. These documents usually correspond to patient documentation and brochures. They show a lower level of expertise: technical terms may be replaced by their non-technical equivalents and be exemplified and defined;

- C_4 or *forum* corpus contains non-expert documents written by patients for patients. This corpus contains messages from the forum indicated above. We expect the corpus to show an even lower level of expertise, although technical terms may also be used.

These corpora are used for the observation and contrastive analysis of selected verbs. C_1/C_4 and C_2/C_3 have comparable sizes.

4.2 Semantic resources

The Snomed International terminology (Côté, 1996) is structured into eleven semantic axes, which we exploit to build the resource that contains the following semantic categories of terms:

- $\mathcal{T}$: 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)*).

Terms from these categories are exploited to semantically annotate our corpora. 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. In relation to FS, we expect these categories to be indicative of frame elements (FEs), while the individual terms should correspond to lexical units (LUs). For instance, the Snomed category *Disorders* should allow us to discover and group under a

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

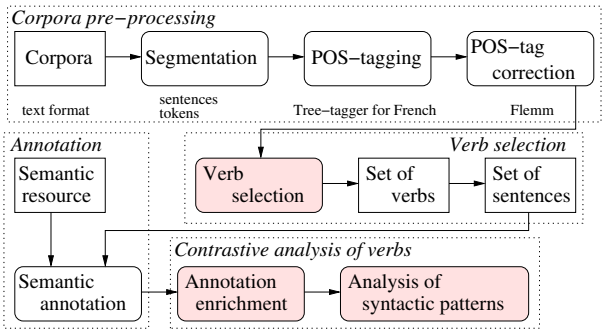

Figure 2: General schema of the method.

single label LUs (*e.g.*, *hypertension* (*hypertension*), *obésité* (*obesity*)) related to the FE DISORDER.

5 Method

The objective is first to discover the descriptions of verbs in a way compatible with FS and then to compare them. The description of verbs depends on the recognition and annotation of noun phrases, such as those provided by the Snomed terminology, which have syntactic dependencies with these verbs. The study is automated as we rely on NLP methods. The proposed method comprises four steps (Figure 2): corpora pre-processing (section 5.1), verb selection (section 5.2), semantic annotation (section 5.3), and contrastive analysis of verbs (section 5.4). On the schema, the three coloured boxes show steps that require human knowledge and that are performed manually; all the other steps are carried out automatically.

5.1 Corpora pre-processing

The corpora are all collected online and properly formatted. They are then tokenized into sentences and words: we expect this may improve POS-tagging. POS-tagging is performed with the French Tree-tagger (Schmid, 1994): its output contains words assigned to parts of speech (*e.g.*, verbs, nouns, adjectives) and lemmatized to their canonical forms (*e.g.*, singular and masculine adjectival forms, infinitive verbal forms). In order to improve the results, we check the output of the POS-tagging with the Flemm tool (Namer, 2000).

5.2 Verb selection

Sets of lemmatized verbs are extracted and their frequencies are computed in the four processed corpora. The verb selection process is carried out according to the following principles:

1. Removing forms that do not correspond to verbs:
   - POS-tagging and lemmatization errors: *e.g.*, *cardiologuer, dolipraner, rhumer*,
   - foreign words, usually also wrongly POS-tagged and lemmatized: *e.g.*, *casemixer, databaser, headacher*,
   - misspellings: *e.g.*, *souaiter, souhiter*.
2. Removing verbs which do not convey a medical meaning (*e.g.*, perception, movement, modal, state verbs);
3. Checking the meaning of the verbs in a medical dictionnary (Manuila et al., 2001): the verbs or their nominal forms have to appear in the dictionnary, as suggested in previous work (Tellier, 2008). For instance, the verb *consulter* is not recorded in the dictionnary but its nominal form *consultation* is: this verb can be then kept at this step;
4. Keeping those verbs with a frequency of 30 occurrences in the corpora. The main corpora considered are C_1 *expert* and C_4 *forum* corpora, while the other two corpora are expected to show at least 10 occurrences of the verbs. As a matter of fact, the frequency indicator is used mainly to guarantee that the verbs have a sufficient number of occurrences and appear in a high number of contexts, these showing a fair level of variability.

After the selection process, we obtain *causer* (*cause*), *traiter* (*treat*), *détecter* (*detect*), *développer* (*develop*), *doser* (*dose*) and *activer* (*activate*) among the remaining verbs. Sentences containing the selected verbs are extracted from each corpus.

5.3 Semantic annotation

The sets of sentences collected at the previous step are annotated using the Ogmios platform (Hamon and Nazarenko, 2008), which integrates and combines several NLP tools. In addition to the syntactic annotation, semantic annotation is obtained after the projection of the semantic resource described in section 4.2: the categories label the participants (that are likely to correspond to FEs), while the specific terms correspond to LUs. Thus, we assume that semantic categories provided by Snomed are useful for the description of semantic frames in medical corpora and that terms from

Step	Number
0. Raw list of verbs	6,218
1. Removing errors and foreign words	3,179
2. Removing non-medical verbs	556
3. Checking the verb meaning	47
4. Checking the frequencies	21

Table 2: Results of the verb selection at each step.

this terminology are useful for the automatic detection of relevant LUs. In a way, our approach is similar to previous work on automatic labeling of semantic roles (Gildea and Jurafsky, 2002; Padó and Pitel, 2007), although in our study we focus on specialized domain material, both corpora and resource, and we have no preconception about the semantic roles associated with medical verbs. Indeed, we exploit the entire Snomed International terminology (except the modifiers).

5.4 Contrastive analysis of verbs

The semantically annotated sentences are then analyzed manually in order to verify if the semantic roles and lexical units are correctly recognized. Wherever necessary, these annotations are enriched manually. This may apply to both missing or unrecognized LUs and FEs. Once the semantic annotation and labeling are completed, verbs from different corpora are analyzed in order to study the differences and similarities which may exist between their uses in these corpora.

6 Results and Discussion

The results are discussed along the following lines: verb selection (section 6.1), semantic annotation (section 6.2), and contrastive analysis of verbs (section 6.3).

6.1 Verb selection

Table 2 indicates the numbers of verbs selected at each step. We can see that an important number of verbs that were removed corresponds to errors, misspellings, and non-medical verbs. The subset of verbs which convey medical meanings corresponds to 0.76% (n=47) of the original set. The final subset contains 21 verbs. From this subset, we selected four verbs for a fine-grained analysis: *observer, détecter, développer*, and *activer*. These verbs were selected for two reasons: they

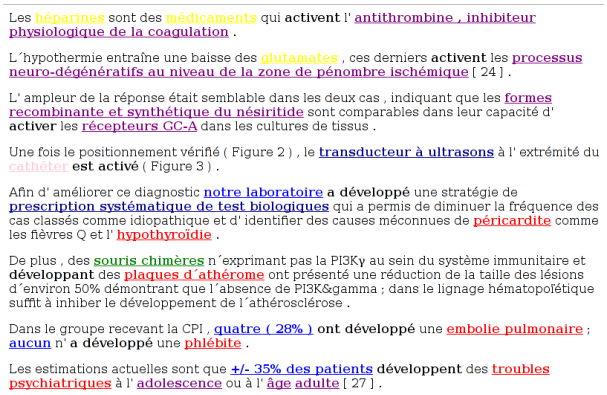

Figure 3: Examples of annotations in C_1. Verbs are in bold characters, semantic labels for arguments with different colours: DISORDERS in red, FUNCTIONS in purple, CHEMICALS in yellow, LIVING ORGANISMS in green, PHYSICAL AGENTS in pink.

were found a high number of contexts (respectively 270, 74, 193 and 85 contexts in C_1 and C_4 corpora) and these contexts seem to be diversified.

6.2 Semantic annotation

Sentences corresponding to the selected verbs have been automatically annotated with semantic classes that are indicative of FEs. The resulting annotation was checked and enriched manually: few errors are detected (*e.g.*, in English-language sentences, *or* (*où in French*) annotated as CHEMICALS (*gold*)). The main limitation is due to the incompleteness of annotations (*facteur* (*factor*) instead of *facteur V de Leiden* (*Factor V Leiden*)) and missing LUs (*e.g.*, *site d'insertion* (*insertion site*) as TOPOGRAPHY, *risque* (*risk*) as FUNCTION, *les traumatisés crâniens* (*people with brain injury*) as LIVING ORGANISMS), usually not recorded in the terminology. An example of the completed annotations is presented in Figure 3. We can observe that these annotations are evocative of those in Figure 1. In Figure 3, the verbs are in bold characters, while different FEs appear in different colours: DISORDERS in red, FUNCTIONS in purple, CHEMICALS in yellow, LIVING ORGANISMS in green, PHYSICAL AGENTS in pink. The syntactic information is also associated with the corresponding LUs but not presented in the figure. The LUs mainly correspond to nouns or noun phrases.

Another limitation discovered at this step is due to the erroneous POS-tagging. For instance, among the 32 contexts of the verb *activer* in C_4, 15 correspond to its adjectival forms (*e.g.*, *j etais une*

Verb	C_1	C_4
observer	$\mathcal{L}, \mathcal{J}, \mathcal{F}, \mathcal{S}, \mathcal{A}, \mathcal{D}$	$\mathcal{L}, \mathcal{J}, \mathcal{F}, \mathcal{A}$
détecter	$\mathcal{L}, \mathcal{A}, \mathcal{J}, \mathcal{P}, \mathcal{F}, \mathcal{D}, \mathcal{T}$	
activer	$\mathcal{C}, \mathcal{F}, \mathcal{P}$	$\mathcal{L}, \mathcal{P}, \mathcal{T}$
développer	$\mathcal{P}, \mathcal{D}, \mathcal{L}, \mathcal{F}$	$\mathcal{L}, \mathcal{D}, \mathcal{F}, \mathcal{T}$

Table 3: The most frequent arguments of verbs.

personne tres active (*I have been a very active person*), *marche active* (*active walking*)). These are not analyzed in the current study. Hence, the resulting number of contexts that were analyzed for this verbs is lower than that of the three other verbs.

6.3 Contrastive analysis of verbs

The contrastive analysis is performed manually. The most frequent labels for FEs of the four verbs analyzed appear in Table 3. We can observe for instance that LIVING ORGANISM $\mathcal{L}$ is usually the most frequent label and appears in both corpora. Typically, it corresponds to human subjects (people communicating in forum discussions in C_4, medical staff and patients observed by the medical staff in C_1). In C_1, PROCEDURES, DISORDERS and CHEMICALS also occupy an important place. Interestingly, with the verb *détecter*, the labels for FEs are identical in both corpora.

Table 4 shows the most frequent patterns of FEs with N_0 (subject) and N_1 (object) functions. We can see that some patterns are common to the two corpora studied (examples (1) to (4)). In the examples presented, the misspellings are genuine.

(1) $\mathcal{P}$ $\mathcal{D}$ with *détecter*: *j'ai acheter un tensiomètre*$_{\mathcal{P}}$ *qui détecte les anomalie cardiaque*$_{\mathcal{D}}$ (*I bought a blood pressure monitor*$_{\mathcal{P}}$ *that detects cardiac abnormality*$_{\mathcal{D}}$)

(2) $\mathcal{J}$ $\mathcal{D}$ with *détecter*: *suite a plusieurs analyses le Medecin*$_{\mathcal{J}}$ *a détecter une péricardite aigüe*$_{\mathcal{D}}$ (*after several tests the Doctor*$_{\mathcal{J}}$ *detected acute pericarditis*$_{\mathcal{D}}$)

(3) $\mathcal{D}$ as N_1 with *développer*: *Un syndrome de détresse respiratoire aiguë*$_{\mathcal{D}}$ *s'est développé* (*Acute respiratory distress syndrome*$_{\mathcal{D}}$ *appeared*)

(4) $\mathcal{D}$ $\mathcal{D}$ with *détecter*: *Une prééclampsie précoce ou sévère*$_{\mathcal{D}}$ *augmente le risque de développer une hypertension chronique*$_{\mathcal{D}}$ *et des maladies cardiovasculaires*$_{\mathcal{D}}$. (*Early or severe pre-eclampsia*$_{\mathcal{D}}$ *increases the risk to develop chronic hypertension*$_{\mathcal{D}}$ *and cardiovacular diseases*$_{\mathcal{D}}$.)

Verb	N_0	N_1	C_1	C_4
observer	$\mathcal{L}$	$\mathcal{D}$	20	3
		$\mathcal{D}$	38	1
	$\mathcal{J}$	$\mathcal{F}$	16	2
	$\mathcal{J}$	$\mathcal{D}$	4	2
détecter	$\mathcal{J}$	$\mathcal{D}$	6	39
	$\mathcal{P}$	$\mathcal{D}$	19	14
	$\mathcal{P}$	$\mathcal{F}$	2	–
	$\mathcal{J}$	$\mathcal{F}$	–	6
	$\mathcal{A}$	$\mathcal{D}$	–	2
activer	$\mathcal{L}$	$\mathcal{P}$	–	3
	$\mathcal{F}$	$\mathcal{T}$	–	2
	$\mathcal{T}$	$\mathcal{F}$	–	1
	$\mathcal{C}$	$\mathcal{F}$	3	–
	$\mathcal{F}$	$\mathcal{F}$	4	–
	$\mathcal{J}$	$\mathcal{J}$	1	–
développer	$\mathcal{L}$	$\mathcal{D}$	12	25
		$\mathcal{P}$	37	–
		$\mathcal{D}$	14	12
	$\mathcal{F}$	$\mathcal{D}$	3	4
	$\mathcal{D}$	$\mathcal{D}$	2	3
		$\mathcal{T}$	–	4

Table 4: The most frequent patterns of arguments of verbs within C_1 and C_4, with their frequencies.

On the other hand, other patterns are specific to a given corpus (examples (5) to (8)).

(5) $\mathcal{T}$ as N_1 with *développer* in C_4: *Certaines personnes réussissent à développer des branches de leurs coronaires*$_{\mathcal{T}}$ (*Some people can develop branches of their coronaries*$_{\mathcal{T}}$)

(6) $\mathcal{P}$ as N_1 with *développer*: in the *expert* corpus, a lot of PROCEDURES (*méthodes de surveillance du foetus (methods for foetus survey), stratégie diagnostique individualisée (strategies for personalized diagnosis), télémédecine (telemedicine)*) are developed with high priority within biomedical research, while this fact is missing in forum discussions

(7) $\mathcal{F}$ $\mathcal{F}$ with *activer* in C_1: *les formes recombinante et synthétique du nésiritide*$_{\mathcal{F}}$

sont comparables dans leur capacité d'activer les récepteurs GC-A$_{\mathcal{F}}$ (*recombinant and synthetic forms of nesiritide*$_{\mathcal{F}}$ *are comparable by their capacity to activate GC-A receptors*$_{\mathcal{F}}$)

(8) $\mathcal{C}$ $\mathcal{F}$ with *activer* in C_1: *Les héparines*$_{\mathcal{C}}$ *sont des médicaments*$_{\mathcal{C}}$ *qui activent l'antithrombine, inhibiteur physiologique de la coagulation*$_{\mathcal{F}}$ (*Heparine*$_{\mathcal{C}}$ *is a medication*$_{\mathcal{C}}$ *that activates antithrombin, physiological inhibitor of the coagulation*$_{\mathcal{F}}$)

Interestingly, the example (5) shows an occurrence of a different meaning of *développer* from that shown in the previous examples. Notice that we have also extracted non-medical meanings of the verbs (examples (9) and (10)), that cannot be labeled with the semantic resource we use.

(9) *Tazzy, tu peux développer ???* (*Tazzy, could you develop???*)

(10) *Santé Canada a développé une nouvelle brochure sur la déclaration des effets indésirables...* (*Health Canada designed a new brochure for the declaration of adverse reactions...*)

More generally, the verb *développer* is used in six patterns common to the two corpora, and eight and five patterns specific to C_1 and C_4 respectively, while the verb *détecter* appears in six common patterns and six specific to each of the corpora. No common pattern was identified for the verb *activer*: the syntactic and semantic properties of this verb are thus different in the two studied corpora, which may also be due to the small set of available contexts. Another difference between these two corpora is that in C_4, we can find some contexts in which verbs do not instantiate all the expected FEs: some syntactic positions remain empty.

On the whole, our observations indicate that the studied verbs present several common patterns within C_1 and C_4. This means that, in this situation, these verbs, although they have a medical meaning, can be correctly understood by patients. When the FEs are partially instantiated, differ from one corpus to the other, or when they show an important difference in terms of frequency, we assume that this may indicate situations in which the understanding may be partial or even unsuccessful. In this case, more thorough explanations are needed by patients to fully understand their health condition and required treatment.

7 Conclusion and Future work

We proposed an NLP approach to automatically discover the participants of verbs and label them using an existing medical terminology assuming that the semantic classes of the terminology are indicative of frame elements (FEs) within the framework of Frame Semantics. The study was performed with medical corpora differentiated according to their levels of expertise: high expertise in C_1 and low in C_4. The contrastive analysis of verbs was done on the basis of automatic annotations completed manually when necessary. The analysis indicates that some verbs share FEs in the studied corpora, while they usually select different FEs according to corpora.

For future work, we plan to add to this study the analysis of C_2 and C_3, which we expect may show intermediate patterns or provide a transition between C_1 and C_4. We also plan to extend this study to other verbs. Up to now, we studied verbal arguments in two syntactic positions (N_0 and N_1), which seems to suffice for the four verbs presented in this paper, but more complex patterns are likely to appear with other verbs. Moreover, automatic distinction between core FEs and non-core FEs (Hadouche et al., 2011), and between the syntactic positions of the labeled entities are other directions for future work.

Our findings may be helpful in several contexts: improving mutual understanding between medical staff and patients, creating two-fold dictionaries with expert and patient expressions, adapting the content of scientific literature for patients. This last context may also provide an interesting application and the possibility for the evaluation of the proposed analysis of verbs.

References

AMA. 1999. Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, 281(6):552–7.

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

R Basili, C Giannone, and D De Cao. 2008. Learning domain-specific framenets from texts. In *ECAI Workshop on Ontology Learning and Population*.

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.

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.

A Condamines. 1993. Un exemple d'utilisation de connaissances de sémantique lexicale: acquisition semi-automatique d'un vocabulaire de spécialité. *Cahiers de lexicologie*, 62:25–65.

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

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

P Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

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

D Gildea and D Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

E Godbert, M Malik, and J Royauté. 2007. Analyse des formes prédicatives dans des textes biomédicaux, pour l'identification d'interactions géniques. In *JOBIM*, pages 81–86.

F Hadouche, S Desgroseilliers, J Pimentel, M.-C. L'Homme, and G Lapalme. 2011. Identification des participants de lexies prédicatives : évaluation en performance et en temps d'un système automatique. In *TIA 2011*.

T Hamon and A Nazarenko. 2008. Le développement d'une plate-forme pour l'annotation spécialisée de documents web: retour d'expérience. *TAL*, 49(2):127–154.

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

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

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

MC L'Homme. 2012. Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, 28:233–252.

L. Manuila, A. Manuila, P. Lewalle, and M. Nicoulin. 2001. *Dictionnaire médical*. Masson, Paris. 9^{e} édition.

A McCray. 2005. Promoting health literacy. *Journal of American Medical Informatics Association*, 12:152–163.

M Miwa, P Thompson, and S Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–65.

F Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523–547.

KH Ohara, 2009. *Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru*, pages 163–182.

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

M Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *GenLex-09*.

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

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

A Roberts, R Gaizauskas, M Hepple, and Y Guo. 2008. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9(11):3–.

CJ Rupp, P Thompson, WJ Black, J McNaught, and S Ananiadou. 2010. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. In *Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features*.

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.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, pages 44–49, Manchester, UK.

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

C Tellier. 2008. Verbes spécialisés en corpus médical: une méthode de description pour la rédaction d'articles terminologiques. Technical report, Université de Montréal.

P Thompson, J McNaught, S Montemagni, N Calzolari, R del Gratta, V Lee, S Marchi, M Monachini, P Pezik, V Quochi, CJ Rupp, Y Sasaki, G Venturi, D Rebholz-Schuhmann, and S Ananiadou. 2011. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12:397.

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.