Frame Semantics-based Study of Verbs across Medical Genres

Ornella WANDJI TCHAMI[a,1], Marie-Claude L'HOMME[b] and Natalia GRABAR[a]
[a] *CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France*
[b] *OLST, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal H3C 3J7, Québec, Canada*

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 achieve 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. These results are important for studying the understanding of medical information by patients and for improving the communication between patients and medical doctors.

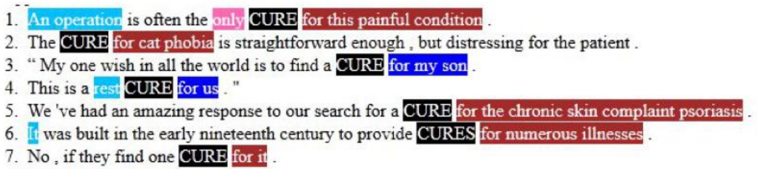
Keywords. Medical informatics, Consumer health informatics, Terminologies, Natural Language Processing, Frame Semantics, France

Introduction

The field of medicine 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 (*eg*, medical doctors, pharmacists, medical students). Moreover, actors with different levels of expertise interact, while their mutual understanding might not always be completely successful. This specifically applies to patients and medical doctors [1,2]. We propose to perform a contrastive and comparative analysis of written medical corpora in French, which are differentiated according to their levels of expertise. More specifically, we concentrate on the study of selected verbs used in these corpora and aim to characterize their syntactic and semantic features. We assume that verbs are an excellent starting point for modeling the semantics of sentences. The description of verbs is based on the Frame Semantics (FS) framework [3]. The FS framework is increasingly used for the description of lexical units in different languages [4-6]. Until recently, French has been neglected with regard

[1] Corresponding Author.

to this framework. Moreover, this framework can be adapted for processing data related to specialized languages [7-9]. FS puts forward the notion of “frames”, which are defined as conceptual scenarios that underlie lexical realizations in language. For instance, in FrameNet [10], the frame CURE (Fig. 1) is described as a frame containing specific Frame Elements (FEs), (such as HEALER, AFFLICTION, PATIENT, TREATMENT), and including lexical units (LUs) such as *cure, alleviate, heal, healer, incurable, treat*.



1. An operation is often the only CURE for this painful condition .
2. The CURE for cat phobia is straightforward enough , but distressing for the patient .
3. “ My one wish in all the world is to find a CURE for my son .
4. This is a rest CURE for us . "
5. We 've had an amazing response to our search for a CURE for the chronic skin complaint psoriasis .
6. It was built in the early nineteenth century to provide CURES for numerous illnesses .
7. No , if they find one CURE for it .

Figure 1. Examples of FrameNet frames for the verb *cure*.

According to our hypothesis, an FS-like modeling should allow us to describe the syntactic and semantic properties of verbs and uncover linguistic differences in corpora of different levels of expertise. Our study relies on the use of terminologies for annotation of verb arguments in corpora differentiated by their expertise levels, on Natural Language Processing (NLP) and on contrastive analysis of these verbs.

Material

Corpora. We study three medical corpora dealing with the specific field of cardiology. These corpora are distinguished according to their levels of expertise and discursive specificities [11]. *Expert* corpus (1,310,115 occurrences) contains expert documents written by medical experts for medical experts (its documents usually correspond to scientific publications, and has a high level of expertise); *student* corpus (395,215 occurrences) contains expert documents written by medical experts for medical students (its documents usually correspond to didactic support created for students, and has a middle level of expertise, in which technical terms are usually introduced and defined); *forum* corpus (1,638,167 occurrences) contains non-expert documents written by patients for patients (it contains messages from the Doctissimo forum *Hypertension Problemes Cardiaques*, and has an even lower level of expertise, although technical terms may also be used).

Semantic resources. We use the Snomed International terminology [12] because it contains an extensive set of terms in French (n=144,267). The terms are structured into several semantic axes, which we also exploit for the semantic annotation of the three corpora: *T* Topography or anatomical locations (*coeur (heart), vaisseau (vessel)*); *S* Social status (*mari (husband), ancien fumeur (former smoker)*); *P* Procedures (*césarienne (caesarean), transducteur à ultrasons (ultrasound transducer)*); *L* Living organisms, such as bacterias and viruses, but also human subjects (*patients, tu (you)*); *J* Professional occupations (*anesthésiste (anesthesiologist)*); *F* Functions of the organism (*pression artérielle (arterial pressure), insuffisance (deficiency)*); *D* Disorders (*obésité (obesity), hypertension artérielle (arterial hypertension)*); *C* Chemical products (*médicament (medication), sodium*); *A* Physical agents (*prothèses (prosthesis), cathéter (catheter)*). 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 single label LUs *(hypertension (hypertension), obésité (obesity))* related to the FE DISORDER.

Methods

Corpora pre-processing. The corpora are collected online and properly formatted. They are then tokenized and POS-tagged with the French Tree-tagger [13]: its output contains words assigned to parts of speech (verbs, nouns, adjectives) and lemmatized to their canonical forms (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 [14].

Verb selection. Sets of lemmatized verbs are extracted and their frequencies are computed. The verb selection relies on the following principles: (1) Remove forms that do not correspond to verbs, such as POS-tagging errors (*cardiologuer, dolipraner, rhumer*), foreign words usually incorrectly lemmatized (*case-mixer, databaser, headacher*), and misspellings; (2) Remove verbs that do not convey a medical meaning (perception, movement, modal, or state verbs); (3) Check the meaning of the verbs in a medical dictionary [15], for which verbs or their nominal forms have to appear there [16]. For instance, the verb *consulter (consult)* is not recorded in the dictionary but its nominal form *consultation* is; (4) Keep those verbs that have a frequency of 30 occurrences in the corpora. Sentences containing the selected verbs are extracted.

Semantic annotation. The sets of sentences are annotated using the Ogmios platform [17]. In addition to the syntactic annotation, semantic annotation is obtained after the projection of the semantic resource: the categories label the arguments (likely to correspond to FEs), while the specific terms correspond to LUs. We assume that categories from Snomed are useful for the description of semantic frames in medical corpora and terms from this terminology are useful for the detection of relevant LUs.

Contrastive analysis of verbs. The semantically annotated sentences are 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, the verbs from different corpora are analyzed in order to study the differences and similarities which may exist between the medical discourse and the uses of verbs in these corpora.

Results and Discussion

During the verb selection, an important number of verbs were removed because they correspond to errors, misspellings, and non-medical meanings. 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 (observe), détecter (detect), développer (develop),* and *activer (activate).* These verbs were selected for two reasons: they were found in a high number of contexts and these contexts seem to be diversified. Sentences corresponding to the selected verbs have been automatically annotated with the semantic resource indicative of FEs. The resulting annotation was checked and enriched manually. We observe that few errors are generated. The main limitations are due to partial annotations *(facteur (factor))* instead of *facteur V de Leiden (Factor V Leiden))* and missing LUs *(site d'insertion (insertion site))* as TOPOGRAPHY, *risque (risk)* as FUNCTION, usually not recorded in the terminology. The syntactic information is also associated with the corresponding LUs (mainly nouns or noun phrases). Another limitation discovered at this step is due to the erroneous POS-tagging. For instance, among the 32 contexts in which the verb *activer* appears in the *forum* corpus, 15 correspond to its adjectival forms (*marche active (active walking)*). The contrastive analysis is performed manually. Among the most frequent labels for FEs, we can observe for instance that LIVING ORGANISM *L* is the most frequent label and appears in all corpora. Typically, it corresponds to human subjects (people communicating in forum discussions in the *forum corpus*, medical staff and patients observed by the medical staff in the *expert* corpus*)*. In the *expert corpus*, PROCEDURES, DISORDERS and CHEMICALS are also very important. Interestingly, with the verb *détecter*, the labels for FEs are identical in both corpora. Among the most frequent patterns of FEs with N0 (subject) and N1 (object) functions, we can observe some common patterns in different corpora (examples below). In the examples, the misspellings are genuine:

P détecter D: *j'ai acheter un tensiomètre$_P$ qui détecte les anomalie cardiaque$_D$*

(I bought a blood pressure monitor$_P$ that detects cardiac abnormality$_D$

D développer D: *Une prééclampsie précoce ou sévère$_D$ augmente le risque de développer une hypertension chronique$_D$ et des maladies cardiovasculaires$_D$.*

(Early or severe pre-eclampsia$_D$ increases the risk to develop chronic hypertension$_D$ and cardiovacular diseases$_D$.)

Other patterns remain specific to a given corpus. More generally, the verb *développer* is used in six patterns common to the two corpora, and eight and five patterns specific to the *expert* and *forum* corpora respectively, while *détecter* appears in six common patterns and six specific to each of the corpora. No common pattern was identified for *activer*. Among the specific patterns, we can find:

développer T (forum corpus): *Certaines personnes réussissent à développer des branches de leurs coronaires$_T$*

développer P (expert corpus): procedures like *méthodes de surveillance du foetus, stratégie diagnostique individualisée, télémédecine* are developed with high priority within biomedical research, while this fact is missing in forum discussions.

F activer F: *les formes recombinante et synthétique du nésiritide$_F$ sont comparables dans leur capacité d'activer les récepteurs GC-A$_T$*

Another difference is that in *forum* corpus, we can find some contexts in which verbs do not instantiate all the expected FEs. The *student* corpus has intermediate position. We assume that when the verbs present common patterns within corpora, 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 show an important difference of frequency, this may indicate that the understanding may be partial and that more thorough explanations are needed for patients. The finding proposed by this study can be used for automatically processing documents designed for patients and for making the content of these documents more appropriate.

Conclusion and Future work

We proposed an NLP approach for discovering the participants of verbs and labelling 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 three medical corpora differentiated according to their levels of expertise: high, low and intermediate. The contrastive analysis of verbs was done on the basis of automatic annotations completed manually when necessary. The analysis indicates that some occurrences of verbs share FEs in the studied corpora, while others select different FEs according to corpora. We plan to extend the study to other verbs and other syntactic positions (in addition to N0 and N1). In addition to the automatic semantic annotation performed in this study, the automatic distinction between core (mandatory) FEs and non-core (optional) FEs [18], and between the syntactic positions of the labeled entities are other directions. Our findings may be helpful for improving understanding between medical staff and patients, and adapting the content of scientific literature for patients. They can also be used for the syntactic simplification of medical documents [19].

References

[1] A McCray. Promoting health literacy. *Journal of American Medical Informatics Association 2005*, 12, 152-163.

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

[3] C Fillmore. *Frame Semantics, 1982,* 111--137.

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

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

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

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

[8] T Schmidt. *The Kicktionary – A Multilingual Lexical Resource of Football Language 2009*, 101-134.

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

[10] J Ruppenhofer, M Ellsworth, MRL Petruck, CR. Johnson, and J Scheffczyk. Framenet II: Extended theory and practice. Technical report, FrameNet, 2006. J Pearson. *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam/Philadelphia, 1998

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

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

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

[14] L. Manuila, A. Manuila, P. Lewalle, and M. Nicoulin. *Dictionnaire médical*. Masson, Paris, 2001.

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

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

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

[18] L Brouwers, L Bernhard D, Ligozat A.and T François. Simplification syntaxique de phrases pour le français. In TALN 2012, 211–224, Montpellier