Unsupervised method for the acquisition of general language paraphrases for medical compounds

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

Thierry Hamon
LIMSI-CNRS, BP133, Orsay
Université Paris 13
Sorbonne Paris Cité, France
hamon@limsi.fr

Abstract

Medical information is widespread in modern society (*e.g.* scientific research, medical blogs, clinical documents, TV and radio broadcast, novels). Moreover, everybody's life may be concerned with medical problems. However, the medical field conveys very specific and often opaque notions (*e.g.*, *myocardial infarction, cholecystectomy, abdominal strangulated hernia, galactose urine*), that are difficult to understand by lay people. We propose an automatic method based on the morphological analysis of terms and on text mining for finding the paraphrases of technical terms. Analysis of the results and their evaluation indicate that we can find correct paraphrases for 343 terms. Depending on the semantics of the terms, error rate of the extractions ranges between 0 and 59%. This kind of resources is useful for several Natural Language Processing applications (*i.e.*, information extraction, text simplification, question and answering).

1 Background

Medical and health information is widespread in the modern society in light of pressing health concerns and of maintaining of healthy lifestyles. Besides, it is also available through modern media: scientific research, articles, medical blogs and fora, clinical documents, TV and radio broadcast, novels, discussion fora, epidemiological alerts, etc. Still, availability of medical and health information does not guarantee its easy and correct understanding by lay people. The medical field conveys indeed very technical notions, such as in example (1).

(1) *myocardial infarction, cholecystectomy, erythredema polyneuropathy, acromegaly, galactosemia*

Although technical, these notions are nevertheless important for patients (AMA, 1999; McCray, 2005; Eysenbach, 2007; Oregon Evidence-based Practice Center, 2008). It has been shown that in several situations such notions cannot be correctly understood by patients: the steps needed for the medication preparing and use (Patel et al., 2002); the instructions on drugs from patient package inserts, and the information delivered in informed consensus and health brochures: it appears that among the 2,600 patients recruited in two hospitals, 26% to 60% cannot manage information available in these sources (Williams et al., 1995); health information in different languages (English, Spanish, French) provided in websites created for patients require high reading level (Berland et al., 2001; Hargrave et al., 2003; Kusec, 2004) and remains difficult to manage by patients, which can be negative for the communication between patients and medical professionals, and the healthcare process (Tran et al., 2009). This situation sets the context of our work. Our objective is to propose method for the automatic acquisition of paraphrases for technical medical notions. More particularly, we propose to concentrate on terms and their words that show neoclassical compounding word formation (Booij, 2010; Iacobini, 1997; Amiot and Dal, 2005), such as in the example (1). Such words often involve Latin and Greek roots or bases, which makes them more difficult to understand, as such words must be decomposed first (see examples (2) and (3)). To our knowledge, this kind of approach has not been applied for the acquisition of laymen paraphrases.

(2) *myocardial* is formed with Latin *myo (muscle)* and Greek *cardia (heart)*

(3) *cholecystectomy* is formed with Greek *chole (bile)*, Latin *cystis (bladder)*, and Greek *ectomy (surgical removal)*

Our work is related to the following research topics:

- *Readability*. The readability studies the ease in which text can be understood. Two kinds of readability measures are distinguished: classical and computational (François, 2011). Classical measures are usually based on number of characters and/or syllables in words, sentences or documents and on linear regression models (Flesch, 1948; Gunning, 1973; Dubay, 2004). Computational measures, that are more recent, can involve vectorial models and a great variety of descriptors. These descriptors, usually specific to the texts processed, are for instance: combination of classical measures with medical terminologies (Kokkinakis and Toporowska Gronostaj, 2006); n-grams of characters (Poprat et al., 2006); discursive descriptors (Goeuriot et al., 2007); lexicon (Miller et al., 2007); morphological descriptors (Chmielik and Grabar, 2011); combination of various descriptors (Wang, 2006; Zeng-Treiler et al., 2007; Leroy et al., 2008; François and Fairon, 2013).

- *Lexical simplification*. The lexical simplification helps to make text easier to understand. Lexical simplification of texts in English has been addressed during the *SemEval* 2012 challenge[a]. Given a short input text and a target word in English, and given several English substitutes for the target word that fit the context, the goal was to rank these substitutes according to how simple they are (Specia et al., 2012). Several clues have been applied: lexicon extracted from oral corpus and Wikipedia, Google n-grams, WordNet (Sinha, 2012); word length, number of syllables, mutual information and frequency of words (Jauhar and Specia, 2012); frequency in Wikipedia, word length, n-grams of characters and of words, syntactic complexity of documents (Johannsen et al., 2012); n-grams, frequency in Wikipedia, n-Google grams (Ligozat et al., 2012); WordNet and word frequency (Amoia and Romanelli, 2012).

- *Dedicated resources*. The building of resources suitable for performing the simplification is another related research topics. Such resources are mainly two-fold lexica in which specialized and non-specialized vocabularies are aligned (in examples (4) to (6), the technical terms are followed by their non-technical equivalents). The first initiative of the kind appeared with the collaborative effort Consumer Health Vocabulary (Zeng and Tse, 2006) (examples in (4)). One of the methods was applied to the most frequently occurring medical queries aligned to the UMLS (Unified Medical Language System) concepts (Lindberg et al., 1993). Another work exploited a small corpus and several statistical association measures for building aligned lexicon with technical terms from the UMLS and their lay equivalents (Elhadad and Sutaria, 2007). Similar work in other languages followed. In French, researchers proposed methods for the acquisition of syntactic variation (Deléger and Zweigenbaum, 2008; Cartoni and Deléger, 2011) from comparable specialized and non-specialized corpora, that led to the detection of verb/noun variations (examples in (5)) and a larger set of syntactic variations (examples in (6)). Besides, research on the acquisition of terminological variation (Hahn et al., 2001), synonymy (Fernández-Silva et al., 2011) and paraphrasing (Max et al., 2012) is also relevant to outline the topics.

(4) {*myocardial infarction*, *heart attack*}, {*abortion*, *termination of pregnancy*}, {*acrodynia*, *pink disease*}

(5) {*consommation régulière*, *consommer de façon régulière*} *(regular use)*, {*gêne à la lecture*, *empêche de lire*} *(reading difficulty)*, {*évolution de l'affection*, *la maladie évolue*} *(evolution of the condition)*

(6) {*retard de cicatrisation*, *retarder la cicatrisation*} *(delay the healing)*, {*apports caloriques*, *apport en calories*} *(calorie supply)*, {*calculer les doses*, *doses sont calculées*} *(calculate the dose)*, {*efficacité est renforcée*, *renforcer son efficacité*} *(improve the efficiency)*

[a]http://www.cs.york.ac.uk/semeval-2012/

Our work is closely related to the building of resources dedicated to the lexical simplification. Our objective is to propose method for paraphrasing the technical medical terms (*i.e.* medical compounds) in expressions that are easier to understand by lay people. This aspect is seldom addressed: we can observe that only some examples in (4) are concerned with the paraphrasing of technical and compound terms (*myocardial infarction*, *acrodynia*). We work with the French data. Contrary to previous work, we do not use comparable corpora with technical and non-technical texts. Instead, we exploit terms from an existing medical terminology and corpora built from social media sources. We assume that this kind of corpora may provide lay people equivalents for technical terms. We also rely on the morphological analysis of technical terms. The expected result is to obtain pairs like {*myocardial*, *heart muscle*} or {*cholecystectomy*, *removal of gall bladder*}. In the following, we start with the presentation of the resources used (section 2), we present then the steps of the methodology (section 3). We describe and discuss the obtained results (section 4) and conclude with some directions for future work (section 5).

2 Resources

2.1 Medical terms

The material processed is issued from the French part of the UMLS. It provides syntactically simple terms that contain one word only (*acrodynia*), and syntactically complex terms that contain more than one word (*myocardial infarction*). Syntactically complex terms are segmented in words. Each term is associated to semantic types. When a given word receives more than one semantic type, a manual post-processing allows to disambiguate it: each word is assigned to one semantic type only. Among the semantic types available, we consider the three most common in the medical practice to which the lay people are the most exposed: Anatomy (616 words): describe human body anatomy (*e.g. abdominopelvic*); Disorders (2,283 words): describe medical problems and their signs (*e.g. infarction, diabetes*); Procedures (1,271 words): describe procedures which may be performed by medical staff to detect or cure disorders (*e.g. cholecystectomy*). In what follows, *word* and *term* can be exchangeable and mean either the graphical unit provided by the segmentation, or the medical notion.

2.2 Corpora

	Wiki	*LesDiab*	*DiabDoct*	*HT*	*Dos*
Number of pages/threads	17,525	6,939	387,435	67,652	8,319
Number of articles/messages	17,525	1,438	22,431	12,588	1,124
Number of words	4,326,880	624,571	35,059,868	6,788,361	836,520

Table 1: Size of the corpora exploited.

We use several corpora collected from the social media sources (their sizes are indicated in Table 1):

1. *Wiki* contains French Wikipedia articles downloaded in February 2014, of which we keep those that are categorized under the medical category *Portail de la médecine*;

2. *LesDiab* is collected from the discussion forum *Les diabétiques*[b] posted between June and July 2013. It is dedicated to diabetes;

3. *DiabDoct* is collected in June 2011 from the discussion forum *Diabète* of Doctissimo[c]

4. *HT* is collected in May 2013 from the discussion forum *Hypertension* of Doctissimo[d]

5. *Dos* is collected in May 2013 from the discussion forum *Douleurs de dos (backache)* of Doctissimo[e]

[b] *http://www.lesdiabetiques.com/modules.php?name=Forums*
[c] *http://forum.doctissimo.fr/sante/diabete/liste_sujet-1.htm*
[d] *http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm*
[e] *http://forum.doctissimo.fr/sante/douleur-dos/liste_sujet-1.htm*

The *Wiki* corpus contains encyclopaedic information on several medical notions from Wikipedia. Thanks to the collaborative writing of the articles, these contain mostly correct information about the topics concerned. Other corpora are collected from the dedicated fora (*e.g.* diabetes or backache). We assume that people involved in these discussions may show low, middle or high degree of knowledge about the disorders and related notions. We expect that all our corpora are written in a simple style and that they contain paraphrases of technical terms. From Table 1, we can observe that the corpora vary in size.

3 Methodology for the automatic acquisition of paraphrases for medical compounds

The methodology is designed for analyzing the neoclassical medical compounds and for searching their non-technical paraphrases in corpora. In our approach, the paraphrases may occur alone, such as *heart muscle*, without being accompanied by their technical compounds (*myocarde*). In this case, we need first to acquire the knowledge needed for their automatic detection. We propose to rely on the morphological analysis of terms. The method is composed of four main steps: the processing of terms, the processing of corpora, the extraction of layman paraphrases for technical terms, and the evaluation of the extractions.

3.1 The processing of medical terms

To reach the morphological information on terms we apply three specific processing:

1. *Morpho-syntactic tagging and lemmatization of terms.* The terms are morpho-syntactically tagged and lemmatized with `TreeTagger` for French (Schmid, 1994). The morpho-syntactic tagging is done in context of the terms. If a given word receives more than one tag, the most frequent one is kept. At this step, we obtain term lemmas with their part-of-speech tags, such as in example (7).

   (7) *myocardique/A (myocardial/A)*, *cholécystectomie/N (cholecystectomy/N)*, *polyneuropathie/N (polyneuropathy/N)*, *acromégalie/N (acromegaly/N)*, *galactosémie/N (galactosemia/N)*

2. *Morphological analysis.* The lemmas are then morphologically analyzed with `DériF` (Namer, 2009). This tool performs the analysis of lemmas in order to detect their morphological structure, to decompose them into their components (bases and affixes), and to semantically analyze their structure. We give some examples of the morphological analysis in (8).

   (8) *myocardique/A*: *[[[myo N*] [carde N*] NOM] ique ADJ]*
   cholécystectomie/N: *[[cholécysto N*] [ectomie N*] NOM]*
   polyneuropathie/N: *[poly [[neur N*] [pathie N*] NOM] NOM]*
   acromégalie/N: *[[acr N*] [mégal N*] ie NOM]*
   galactosémie/N: *[[galactose NOM] [ém N*] ie NOM]*

   The computed bases and affixes are associated with syntactic categories (*NOM, ADJ, V*). When a given base is suppletive (does not exist in modern French but was borrowed from Latin or Greek languages), `DériF` assigns the most probable category (*e.g. N** for nouns, *A** for adjectives). For instance, the analysis of *myocardique/A* indicates that this word contains the suppletive noun bases *myo N* (muscle)* and *carde N* (heart)*, and the affix *-ique/ADJ*. We can observe that some bases can be decomposed further (*e.g. galactose* in *galact (milk)* and *ose (sugars)*, *cholecystectomy* in *chole (bile)* and *cystis (bladder)*). The words that contain more than one base are considered to be compounds and are processed in the further steps of the method.

3. *Association of morphological components with French words.* The bases are "translated" with words from modern French. We use for this resource built in previous work (Zweigenbaum and Grabar, 2003; Namer, 2003) (see some examples in (9)).

   (9) *myocardique/A*: *myo=muscle (muscle)*, *carde=coeur (heart)*
   cholécystectomie/N: *cholécysto=vésicule biliaire (gall bladder)*, *ectomie=ablation (removal)*

polyneuropathie/N: *poly=nombreux (several)*, *neuro=nerf (nerve)*, *pathie=maladie (disorder)*
acromégalie/N: *acr=extrémité (extremity)*, *mégal=grandeur (size)*
galactosémie/N: *galactose=galactose (galactose)*, *ém=sang (blood)*

Some words can remain technical (*e.g.*, *galactose*, *vésicule biliaire*), while other components totally lose their technical meaning (*e.g. mégal=grandeur (size)*, *poly=nombreux (several)*).

3.2 The processing of corpora

The corpora are first segmented in words and sentences. Then, we also perform morpho-syntactic tagging and lemmatization with `TreeTagger` for French.

3.3 The extraction of layman paraphrases corresponding to technical terms

French words corresponding to the morphological decomposition of terms (examples in (9)) are projected on corpora in order to extract sentences and their segments which can provide the layman paraphrases for the corresponding technical terms. Sentences that contain the translated French words are extracted as candidates for proposing the paraphrases. Additionally, the segments delimited by these words are also extracted. We consider the co-occurrence of the words issued from the morphological decomposition in a sliding graphical window of n words. In the experiments presented, the window size n is fixed to 10 words. Smaller or larger windows show less performance.

(10) *Les causes de tachycardie ventriculaire sont superposables à celles des extrasystoles ventriculaires: infarctus du myocarde, insuffisance cardiaque, hypertrophie du muscle du coeur et prolapsus de la valve mitrale.*

The sentence in (10) contains words *muscle* and *coeur*, underlined in the example, that correspond to the morphological components of *myocardique* (see examples in (9)). For this reason, this sentence is extracted, as well as the segment delimited by these two words *muscle du coeur (heart muscle)*.

3.4 The evaluation

The objective of the evaluation is to assess whether the proposed method is valid for the acquisition of paraphrases for technical medical terms. The obtained results are evaluated manually by a computer scientist with no training in biomedicine, but with background in computational linguistics and morphology. We analyze the candidates for paraphrases from several points of view: Are the French words corresponding to the components extracted correctly? Do these French words provide valid candidates for paraphrases? How easy are these paraphrases to be understood by laymen or by non-experts in medicine? During the evaluation related to the second point (*Do these French words provide valid candidates for paraphrases?*), we distinguish four situations:

1. the extraction is correct: *e.g. myocardique* paraphrased in *muscle du coeur (heart muscle)*;
2. the extraction suffers from the incorrect morphological decomposition or from the wrong "translation" in French: *e.g. périanal* is "translated" in *autour (around)* and *an (meaning year as it is)*. The "translation" of this last word *an* is not correct and should be *anus (anus)* instead. Because of the wrong "translation", we collect a lot of incorrect segments like *autour de 30 ans (around 30 years)*;
3. the extraction should be post-processed but contains the correct paraphrase: *e.g. spondylarthrose*, "translated" in *vertèbre (vertebra)* and *arthrose (arthrosis)*, is paraphrased in *arthrose que l'on ne voyait pas sur la vertèbre (arthrosis that was not seen on the vertebra)*, while the correct paraphrase from this segment should be *arthrose sur la vertèbre (arthrosis on the vertebra)*;
4. the extraction is wrong and can provide no useful information.

This evaluation allows to estimate precision of the results in three versions: strong precision P_{strong} (only the correct extractions are considered (extractions from 1)); weak precision P_{weak} (correct extractions and extractions that need post-processing are considered (extractions from 1 and 3)); rate of incorrect extractions $\%_{incorrect}$ (the percentage of the incorrect extractions is computed (extractions from 4)).

4 Results and Discussion

4.1 The morphological analysis of terms

We generate the morphological analysis for 218 single words from the anatomy semantic type, 1,789 disorder words and 1,023 procedure words: over 70% of words are morphologically analyzed. Among these words, we observe compounds (*myocardique*) and words formed with affixes (*e.g. réadaptation* derived from *adaptation*, derived in its turn from *adapter*). The remaining words may be simple (*e.g. abcès (abscess)*, *lèpre (leprosy)*, *cicatrice (scar)*) or contain bases and affixes that are not managed by `Dérif` (*e.g. pneumostrongylose (pneumostrongylosis)*, *lagophtalmie (lagophthalmos)*, *nécatorose (necatorosis)*). Among the generated decompositions by `Dérif`, we can find some cases with ambiguous decomposition that occur when medical terms can be decomposed in several possible ways, among which only one is semantically correct. For instance, *posturographie (posturography)* is decomposed into: *[post [[uro N*] [graphie N*] NOM] NOM]*, which may be glossed as *control during the period which follows the therapy done on the urinary system*. From the formal point of view, such decomposition is very possible, although it is weak semantically. For the term *posturographie*, the right decomposition is: *[[posturo N*] [graphie N*] NOM]*, which is related to the *definition of the optimal body position when walking or sitting*. As indicated above, some terms (*e.g. périanal*) can be incorrectly "translated" in French.

4.2 The preprocessing of corpora

Our main difficulty at this step is related to the processing of forum messages and to their segmentation into sentences. In addition to possible and frequent spelling and grammatical errors, forum messages have also a very specific punctuation, which may be missing or convey personal feelings and emotions. This seriously impedes the possibility to provide the correct segmentation in sentences, and means that, because of the missing punctuation, the mapping of decomposed terms with corpora may be done with bigger text segments in which the semantic relations between the mapped components may be weak or non-existent, and provide incorrect extractions. We plan to combine the current method with the syntactic analysis in order to ensure that stronger syntactic and semantic relations exist between the components.

4.3 The extraction of paraphrases and their evaluation

We present the results on extraction of sentences and paraphrases from the corpora processed. In Table 2, for the three semantic types of terms (anatomy *ana.*, disorders *dis.*, and procedures *pro.*) from each corpus, we indicate the following information: the number of different sentences extracted (*sentences*), yje number of different terms (*uniq. terms*), the number of correct paraphrases (*correct*), the number of paraphrases that are possibly correct (*pos. correct*), the number of paraphrases which morphological analysis and "translation" should be improved (*morph. ana.*), and the number of incorrect paraphrases (*incorrect*). The last three lines indicate the precision values: strong precision (P_{strong}), weak precision (P_{weak}) and incorrect extractions ($\%_{incorrect}$).

Number of	*Wiki*			*LesDiab*			*DiabDoct*			*HT*			*Dos*		
	ana	*dis*	*pro*	*ana*	*dis*	*pro*	*ana*	*dis*	*pro*	*ana*	*dis*	*pro*	*ana*	*dis*	*pro*
sentences	1238	4003	999	15	71	10	721	2901	564	246	1233	678	42	708	30
uniq. terms	93	382	154	7	30	5	35	204	48	29	133	42	13	44	13
correct	469	1571	364	3	32	4	227	1189	67	114	637	38	12	466	13
pos. correct	270	868	93	3	7	-	40	332	5	10	85	9	3	98	2
morph. ana.	41	155	323	1	2	6	100	3	394	22	-	591	2	1	12
incorrect	462	1424	220	8	30	-	354	1	98	100	511	40	25	135	3
P_{strong}	38	39	36	20	45	40	32	40	12	46	52	6	29	66	43
P_{weak}	60	61	46	40	55	40	37	52	13	50	59	7	36	80	50
$\%_{incorrect}$	40	39	54	53	42	0	49	47	17	41	41	41	59	20	10

Table 2: Results on the paraphrases extracted and evaluated.

From the data presented in Table 2, we can propose several observations: (1) the *Wiki* corpus, that is not the largest in our dataset, provides the largest number of extractions (sentences and unique terms); (2) among the three semantic types (anatomy, disorders and procedures), the number of paraphrases extracted for disorders is the largest in all corpora; (3) the largest set of paraphrases, that suffer from the incorrect morphological decomposition or "translation", is obtained for the procedure terms. According to these observations, P_{strong} ranges between 20 to 46% for anatomy, 39 and 66% for disorders, and 6 to 43 for procedures. The P_{weak} values, that takes into account the paraphrases that need post-processing, show the increase by 0 to 28% by comparison with the P_{strong} values. The $\%_{incorrect}$ values indicate that anatomy terms show the largest rate (40 to 59%) of incorrect paraphrases: it is possible that the anatomy terms present the lowest rate of compositionality. The incorrect paraphrases are between 20 and 47 among the disorder terms, and between 0 to 54 among the procedure terms. The syntactic analysis may help to improve the current results. On the whole, the proposed method allows to extract the paraphrases for 722 different terms from the corpora processed. Within the evaluated set of extractions, these paraphrases are correct for 273 terms; while 343 terms are provided with correct paraphrases and paraphrases that need to be post-processed. Most of the extracted paraphrases are noun phrases, and, at a lesser extent, verb phrases. We present some examples of the correct paraphrases extracted:

- *dorsalgie (dorsalgia)*: *douleur dans le dos (pain in the back)*
- *myélocyte (myelocyte)*: *cellules dans la moelle osseuse (cells of the bone marrow)*
- *lombalgie (lombalgia)*: *douleurs dans les reins (pain in kidney)*
- *gastralgie (gastralgia)*: *douleurs à l'estomac (stomach pain)*
- *desmorrhexie (desmorrhexia)*: *rupture des ligaments (ligamentous rupture)*
- *hépatite (hepatitis)*: *inflammation du foie (liver inflammation)*

We can find several types of paraphrases that suffer from incorrect decomposition or "translation":

- *syringomyélie (syringomyelia)* is currently "translated" in *moelle (marrow or spinal cord)* and *canal (canal)*. This term means a disorder in which a cyst or cavity forms within the spinal cord. We assume that a more correct "translation" of this term should be: *moelle (marrow or spinal cord)* and *cavité (cavity)*;

- *sous-dural* is "translated" in *sous (sub)* and *dur (hard)*. The term is related to specific space in brain that can be opened by the separation of the arachnoid mater from the dura mater. Concerning its "translation", we assume that *dure-mère (dura mater)* should be used instead of *dur (hard)*. Besides, the names of anatomical locations often remains difficult to understand. We assume that even when terms are decomposed and "translated" correctly, the paraphrases for such terms may be not suitable for laymen: other types of explanations (*e.g.* schemes or pictures) should be used instead;

- *hyperémie (hyperaemia)* is "translated" in *hyper* and *sang (blood)*. The term means the increase of blood flow to different tissues in the body. This term is not fully compositional because the notion of tissues is absent, while necessary for its understanding. The proposed extractions for this term mainly come from corpora related to diabetes, in which *hyper* and *hypo* are often used in relation with the *hyperglycemia* or *hypoglycemia*. This means that *hyper* should be "translated" with other words, such as *increase* or *elevated*;

- *hétérotopie* is translated in *autre (another)* and *endroit (place)*. The term means the displacement of an organ from its normal position and that [an organ] is found in another place than the one expected. This term brings no correct candidates for paraphrases because: it is not fully compositional and its "translation" provides very common words widely used in the corpora.

Among the incorrect extractions we can find: (1) more terms with non-compositional semantics (such as *ostéodermie (osteoderm)*, *causalgie (causalgia)*, *adénoïde (adenoid)*, or *xanthochromie (xanthochromia)*) for which the extracted paraphrases capture only part of the meaning; and (2) extractions that must be controled by the syntactic analysis (*e.g. petite boule de peau qui a sortie entre l'ongle et... (small skinball that appeared between the nail and...)* for *micronychie (micronychia)*) to make them more grammatical. Paraphrases extracted from the *Wiki* corpus cover larger range of medical terms, while those extracted from

fora dedicated to a given medical topics are redundant. On the whole, we can consider that the currently proposed method allows extracting interesting candidates as the paraphrases of technical terms, that are indeed much easier to understand than the technical terms by themselves.

If we compare the obtained results with those presented in previous work, we can observe that:

- we extract paraphrases for larger number of terms: 343 terms with correct and possibly correct paraphrases (722 terms with paraphrases in total) in our work against a total of 65 and 82 in (Deléger and Zweigenbaum, 2008), 109 in (Cartoni and Deléger, 2011), and 152 in (Elhadad and Sutaria, 2007). In our work, the terms may receive more than one paraphrase;
- the precision values we obtain are comparable with those indicated in previous work: 67% and 60% in (Deléger and Zweigenbaum, 2008), 66% in (Cartoni and Deléger, 2011), and 58% in (Elhadad and Sutaria, 2007);
- in the cited work, the content of the corpora is explored but no reference is done to the set of terms expected to be found. Because we work with a termset, we can compute the recall. If we consider the terms that can be analyzed morphologically (3,030 terms), and for which we can find the paraphrases with the proposed method, the recall value is close to 10% with the correct paraphrases (299 terms), and to 24% with all the paraphrases extracted (722 terms). Yet, it is not sure that all of the terms, that have been analyzed morphologically, can be provided with paraphrases in the corpora processed.

Besides, we should not forget that the nature of compounds and the decomposition of terms into components also mean that specific semantic relations exist between these components (Namer and Zweigenbaum, 2004; Booij, 2010). These are inherent to the syntactic constructions extracted. The characteristics of these relations will be described and modeled in future work.

5 Conclusions and Future work

We propose to exploit social media texts in order to detect paraphrases for technical medical terms, concentrating particularly on neoclassical compounds (*e.g.*, *myocardial, cholecystectomy, galactose, acromegaly*). The work is done in French. The method relies on the morphological analysis of terms, on the "translation" of the components of terms in modern French words (*e.g.* {*card*, *heart*}), and on the projection of these words on corpora. The method allows extracting correct and possibly correct paraphrases for up to 343 technical terms. For covering larger set of terms, additional corpora must be treated. The extracted paraphrases are easier to understand than the original technical terms. Moreover, the semantic relations among the components, although non explicated, are conveyed by the paraphrases. We can consider that the method proves to be efficient and promising for the creation of lexicon suitable for the simplification of medical texts. Besides, the purpose of the method is to cover neoclassical compound terms that are usually non treated with automatic approaches, as they do not present clear formal similarity with their paraphrases. One of the difficulties we have currently is related to the lack of constrains on the extracted segments. In future work, we plan to apply the syntactic analysis for parsing the extracted sentences. Another possibility is to compute the probability for a given paraphrase to be correct, which can rely for instance on frequency of the extracted paraphrases, on their syntactic structure, etc. In order to make the extraction of paraphrases more exhaustive, we will apply the method to other corpora and we will use additional resources (synonyms, associative resources) for performing the approximate mapping of paraphrases. In future work, we will take into account syntactically complex terms and not only simple words. The very objective of our work is to exploit and test the resource created for the simplification of medical texts.

Acknowledgments

The authors acknowledge the support of the Université Paris 13 (project BQR Bonus Quality Research, 2011), the support of the MESHS Lille projet Émergent CoMeTe, and the support of the French Agence Nationale de la Recherche (ANR) and the DGA, under the Tecsan grant ANR-11-TECS-012.

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.

D Amiot and G Dal. 2005. Integrating combining forms into a lexeme-based morphology. In *Mediterranean Morphology Meeting (MMM5)*, pages 323–336.

M Amoia and M Romanelli. 2012. Sb: mmsystem - using decompositional semantics for lexical simplification. In **SEM 2012*, pages 482–486, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

GK Berland, MN Elliott, LS Morales, JI Algazy, RL Kravitz, MS Broder, DE Kanouse, JA Munoz, JA Puyol, M Lara, KE Watkins, H Yang, and EA McGlynn. 2001. Health information on the internet. accessibility, quality, and readability in english ans spanish. *JAMA*, 285(20):2612–2621.

Geert Booij. 2010. *Construction Morphology*. Oxford University Press, Oxford.

B Cartoni and L Deléger. 2011. Dcouverte de patrons paraphrastiques en corpus comparable: une approche base sur les n-grammes. In *TALN*.

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.

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

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*.

N Elhadad and K Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *BioNLP*, pages 49–56.

Gunther Eysenbach. 2007. Poverty, human development, and the role of ehealth. *J Med Internet Res*, 9(4):e34.

S Fernández-Silva, J Freixa, and MT Cabré. 2011. A proposed method for analysing the dynamics of cognition through term variation. *Terminology*, 17(1):49–73.

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.

T François. 2011. *Les apports du traitements automatique du langage la lisibilit du franais langue trangre*. Phd thesis, Universit Catholique de Louvain, Louvain.

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

R Gunning. 1973. *The art of clear writing*. McGraw Hill, New York, NY.

Udo Hahn, Martin Honeck, Michael Piotrowsky, and Stefan Schulz. 2001. Subword segmentation - leveling out morphological variations for medical document retrieval. In *AMIA*, 229-33.

Darren Hargrave, Ute Bartels, Loretta Lau, Carlos Esquembre, and Éric Bouffet. 2003. évaluation de la qualité de l'information médicale francophone accessible au public sur internet : application aux tumeurs cérébrales de l'enfant. *Bulletin du Cancer*, 90(7):650–5.

C Iacobini. 1997. Distinguishing derivational prefixes from initial combining forms. In *First mediterranean conference of morphology*, Mytilene, Island of Lesbos, Greece, septembre.

SK Jauhar and L Specia. 2012. Uow-shef: Simplex – lexical simplicity ranking based on contextual and psycholinguistic features. In **SEM 2012*, pages 477–481, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

A Johannsen, H Martínez, S Klerke, and A Søgaard. 2012. Emnlp@cph: Is frequency all there is to simplicity? In **SEM 2012*, pages 408–412, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

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.

Sanja Kusec. 2004. Les sites web relatifs au diabète, sont-ils lisibles ? *Dibète et société*, 49(3):46–48.

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.

AL Ligozat, C Grouin, A Garcia-Fernandez, and D Bernhard. 2012. Annlor: A naïve notation-system for lexical outputs ranking. In **SEM 2012*, pages 487–492.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods Inf Med*, 32(4):281–291.

Aurélien Max, Houda Bouamor, and Anne Vilnat. 2012. Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *EMNLP*, pages 721–31.

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.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, San-Francisco.

F Namer. 2003. Automatiser l'analyse morpho-sémantique non affixale: le système DériF. *Cahiers de Grammaire*, 28:31–48.

F Namer. 2009. *Morphologie, Lexique et TAL : l'analyseur DériF. TIC et Sciences cognitives*. Hermes Sciences Publishing, London.

Oregon Evidence-based Practice Center. 2008. Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved. Technical report, Agency for healthcare research and quality.

V Patel, T Branch, and J Arocha. 2002. Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, 65(3):193–211.

M Poprat, K Markó, and U Hahn. 2006. A language classifier that automatically divides medical documents for experts and health care consumers. In *MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics*, pages 503–508, Maastricht.

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

R Sinha. 2012. Unt-simprank: Systems for lexical simplification ranking. In **SEM 2012*, pages 493–496, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

L Specia, SK Jauhar, and R Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In **SEM 2012*, pages 347–355.

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

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

MV Williams, RM Parker, DW Baker, NS Parikh, K Pitkin, WC Coates, and JR Nurss. 1995. Inadequate functional health literacy among patients at two public hospitals. *JAMA*, 274(21):1677–82.

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

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.

Pierre Zweigenbaum and Natalia Grabar. 2003. Corpus-based associations provide additional morphological variants to medical terminologies. In *AMIA*.