Automatic Extraction of Layman Names for Technical Medical Terms

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

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

***Abstract*—Medical and health information is widespread in the modern society in light of pressing health concerns and of maintaining of healthy lifestyles. It is also available through modern media (scientific research, medical blogs, clinical documents, TV and radio broadcast, novels, etc.) However, medical area conveys very specific and often opaque notions (*e.g.*, *myocardial infarction, cholecystectomy, abdominal strangulated hernia, galactose urine*), which are difficult to understand by people without medical training. We propose an automatic method for the acquisition of paraphrases for technical medical terms. We expect that the paraphrases are easier to understand than the original terms. The method is based on the morphological analysis of terms and on text mining of social media texts. Analysis of the results and their evaluation indicate that such paraphrases can indeed be found in non specialized documents and show easier understanding level. Depending on the semantics of the terms, the precision values of the extractions ranges between 6 and 100%. This kind of resources is useful for several Natural Language Processing applications (*i.e.*, information retrieval and extraction, text simplification and health literacy, question and answering).**

I. 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. The availability of medical and health information is not enough to easily provide people without medical training a correct understanding. If the understanding is acquired and almost natural for medical staff (*e.g.*, medical doctors and students, nurses, pharmacists), ordinary citizens may have some difficulties to understand and to handle it. Very technical notions are indeed widely used in the medical field, such as those presented in example (1).

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

The understanding of these notions is nevertheless important for patients [1], [2], [3], [4], as it has been shown that they play crucial role for successful healthcare process. Yet, it has been also shown that such notions cannot be correctly understood by patients in several situations:

- the steps needed for the medication preparing and use are not fully understood [5];
- the instructions on drugs from patient package inserts, as well as the information delivered for patients in informed consensus and health brochures are not fully understood. For instance, it appears that among the 2,600 patients recruited in two hospitals, 26% to 60% of them cannot understand and manage the health information available in these sources [6];
- health information in different languages (English, Spanish, French) provided in websites created for patients also requires high reading level [7], [8], [9] and remains difficult to manage by patients, which can be negative for the communication between patients and medical professionals and the healthcare process of patients [10].

This situation provides the main motivation for 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 [11], [12], [13], such as in example (1). One particularity is that such words often involve Latin and Greek roots or bases, which makes them even more difficult to understand. Indeed, before it comes to understanding, the user must first to decompose the words and to make the link with the common language. The examples of such decompositions are presented in (2) and (3).

(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)*

In the following of this paper, we start with the presentation of the existing related work (section II) and indicate the objectives of our work (section III). We then present material used (section IV), and the steps of the methodology (section V). We describe and discuss the obtained results (sections VI and VII) and conclude with some directions for future work (section VIII).

II. Related work

Two big research areas are concerned with the work we propose:

1) The Semantic Interoperability guarantees the link between two registers, that are specialized and non-specialized health registers in our work. It analyzes the situation through the study and creation of relations between expert and non-expert languages in order to improve the communication among these two communities. Notice that usually the Semantic Interoperability is involved in the communication between two automatic systems or in the alignment of two terminological systems. Our acceptation of this notion is positioned on different dimension (language register).
2) The interaction between Computer Sciences and Disability is relevant to our work because Computer Sciences, and more particularly Natural Language Processing (NLP), provide methods for helping patients with understanding the health information [14]. In this framework, the disability is concerned with the non-understanding of the technical medical notions.

From the point of view of the NLP area, our work is closely related to several research topics: readability (section II-A), lexical simplification (section II-B), building of dedicated resources (section II-C) and decomposition of neoclassical compounds (section II-D). Notice that these topics are interlinked among them and, taken together, form a sophisticated research problem.

A. Readability

The readability studies the ease in which text can be understood. Two kinds of readability measures are distinguished: classical and computational [15]. Classical measures are usually based on number of characters and/or syllables in words, sentences or documents and on linear regression models [16], [17], [18]. Computational measures, that are more recent, can involve vectorial models and a great variety of descriptors. These descriptors are usually specific to the texts processed. Because medical texts show some specificities [19], [20], the descriptors used are for instance: combination of classical measures with medical terminologies [21]; n-grams of characters [22]; discursive descriptors [23]; lexicon [24]; morphological descriptors [25]; combination of various descriptors [26], [27], [28], [29]. When it comes to the readability of medical and health documents, the *health literacy* term is usually used [19], [20]. The objective of health literacy is to assess the ease with which the health documents and information are understood by patients. A specific case of the health literacy is the *health numeracy*, that addresses understanding and processing of numerical values and information (*e.g.* medicine preparation, dosage and intake).

B. Lexical simplification

The lexical simplification helps to make text easier to understand. Lexical simplification of general language 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 [30]. Several clues have been applied: lexicon extracted from oral corpus and Wikipedia, Google n-grams, WordNet [31]; word length, number of syllables, mutual information and frequency of words [32]; frequency in Wikipedia, word length, n-grams of characters and of words, syntactic complexity of documents [33]; n-grams, frequency in Wikipedia, n-Google grams [34]; WordNet and word frequency [35]. The features related to the frequency of words appear to be among the most efficient for this task.

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

C. 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. For instance, in examples (4) to (6), technical terms are followed by their non technical equivalents. The first initiative of the kind appeared with the collaborative effort Consumer Health Vocabulary [36] (examples in (4)). One of the methods explored was applied to the most frequently occurring medical queries that have been aligned to the UMLS (Unified Medical Language System) concepts [37]. Then, two reviewers assessed the accuracy of this alignment. 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 [38]. Similar work in other languages followed. In French, researchers proposed methods for the acquisition of syntactic variations [39], [40] from comparable specialized and non-specialized corpora, that lead to the detection of equivalent phrases: verb/noun variations (examples in (5)) and a larger set of syntactic variations (examples in (6)). Another work proposed English and French patient-oriented terminology for the breast cancer area [41]. No relations were established with medical terminologies in several existing experiences [39], [40], [41]. Besides, the research on the acquisition of terminological variation [42], synonymy [43] and paraphrasing [44] is also relevant to outline this research 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)*

D. Decomposition of neoclassical compounds

The purpose of the decomposition of neoclassical compounds is to detect the morphological components within these compounds. Some researchers aim at this kind of decomposition in order to improve the information retrieval and indexing results [45], [46], [42]. Indeed, for a term like *iridochoroiditis*, it may be useful to know and to indicate explicitely its components (*inflammation, iris* and *choroid*) for finding additional releval documents. In this way, when a document that does

not contain the compound term but that contains at least one of its components, it can be retrieved nevertheless. In other approaches, that go above the decomposition into components, the semantic relations between these components are also detected and may be described. Several of such studies have been done manually [47], [48], [49], with the objective to describe the semantic relations among the components and to shows that the semantic patterns involved are stable and productive in the medical language. For instance, within the compound *iridochoroiditis* we can observe the *localization* relation, as the *inflammation* is located in *iris* and *choroid*. The automatisation of the process in English and French is proposed [50], [51].

III. Objectives

Our work is closely related to the decomposition of medical neoclassical compounds (section II-D) and the building of resources dedicated to the lexical simplification (section II-C). 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 in the existing work: we can observe that only some examples in (4) are concerned with the paraphrasing of technical and compound terms (*i.e.* {*myocardial infarction*, *heart attack*}, {*acrodynia*, *pink disease*}). We work with the French data. Contrary to the existing work on the detection of paraphrases, 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 this kind of corpora may provide lay people equivalents for technical terms. We also rely on the morphological decomposition and analysis of technical terms. The expected result is to obtain pairs like {*myocardial*, *heart muscle*}, {*cholecystectomy*, *removal of gall bladder*}.

IV. Material

We use three kinds of material: the medical terms for which we look for paraphrases (section IV-A), the corpora from which the paraphrases are extracted (section IV-B), and the linguistic resource which helps to establish the link between the terms and the expressions in corpora (section IV-C).

A. Medical terms

The medical terms processed is issued from the French part of the UMLS. In the UMLS, we can find syntactically simple terms that contain one word only (*e.g. acrodynia*), and syntactically complex terms that contain more than one word (*e.g. myocardial infarction*). Syntactically complex terms are segmented into words.

The UMLS assigns to each term semantic types (*e.g.*, *Anatomical Structure, Organism, Substance, Finding, Health Care Activity, Occupation or Discipline, Language, Idea or Concept*). We use terms from three semantic types (*Anatomical Structure, Finding and Health Care Activity*). We assume indeed that these semantic types are the most frequent in the medical practice and that patients are the most exposed to them. There are some examples of the corresponding words:

- Anatomy or Anatomical Structure (616 words): these words describe human body anatomy (*e.g. abdominopelvic*);
- Finding or Disorders (2,283 words): these words describe medical problems and their signs (*e.g. infarction, diabetes*);
- Health Care Activity or Procedures (1,271 words): these words describe procedures which may be performed by medical staff to detect or cure disorders (*e.g. cholecystectomy*).

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. In what follows, *word* and *term* can be exchangeable and mean either the graphical unit provided by the segmentation of terms, or a given medical notion.

B. Corpora

	Number of threads	*Number of messages*	*Number of words*
LesDiab	1,438	6,939	624,571
DiabDoct	22,431	387,435	35,059,868
HT	12,588	67,652	6,788,361
Dos	1,124	8,319	836,520

TABLE I. Size of the corpora exploited.

We use several corpora collected from the social media sources. The sizes of corpora are indicated in Table I in terms of the number of threads (or discussion topics), the number of messages exchanged between the users, and the number of words the corpora contain. The four corpora used are:

1) *LesDiab* is collected from the discussion forum *Les diabétiques*[b] posted between June and July 2013. It is dedicated to diabetes;
2) *DiabDoct* is collected in June 2011 from the discussion forum *Diabète* of Doctissimo[c]
3) *HT* is collected in May 2013 from the discussion forum *Hypertension* of Doctissimo[d]
4) *Dos* is collected in May 2013 from the discussion forum *Douleurs de dos (backache)* of Doctissimo[e]

We can observe that the corpora are collected from the fora dedicated to a given health condition (*i.e.* diabetes, backache, hypertension). We assume that people involved in the forum discussions may show low, middle or high degree of knowledge about the disorders and related notions. According to the specificities of fora, we expect that these texts are written in a simple and understandable style. These should also contain paraphrases of technical terms, such as they are provided by the contributors, with expressions understandable by laymen. From Table I, we can observe that corpora have very different sizes. The largest is the *DiabDoct* corpus with over 35 M words.

[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*

C. Linguistic resource

The linguistic resource used contains pairs of words {*suppletive base*, *French word*} and allow making the link between the Latin and Greek roots or words with the modern French words. This resource has been built in previous studies [52], [51] and is not specifically dedicated to the currently proposed experiments, although it is dedicated to the specificity of the material processed. The resource contains pairs such as

{*andr*, *male*}, {*ectomie*, *removal*}, {*myo*, *muscle*}
{*para*, *against*}, {*peri*, *around*}

The resource provides 964 {*suppletive base*, *French word*} pairs.

V. Methodology for the automatic acquisition of paraphrases for medical compounds

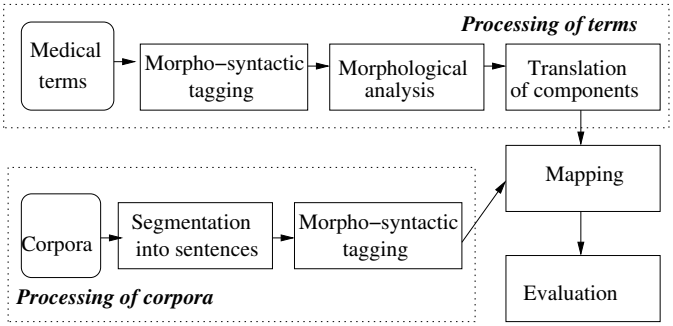

Fig. 1. Methodology for the extraction of lay man paraphrases for medical compounds from corpora.

The methodology is designed for analyzing the neoclassical medical compounds and for searching their non technical paraphrases in corpora. The paraphrases may occur alone, such as *heart muscle*, without being accompanied by their technical compounds (*myocardial*). In this case, we need first to acquire the knowledge needed for their automatic detection. We propose to rely on the morphological analysis of the technical terms. The method is composed of four main steps (Figure 1): processing of terms (section V-A), processing of corpora (section V-B), extraction of layman paraphrases for technical terms (section V-C), and evaluation of the extractions (V-D).

A. Processing of medical terms

In order to reach the morphological information on terms, we apply three specific treatments:

1) *Morpho-syntactic tagging and lemmatization of terms.* The terms are morpho-syntactically tagged and lemmatized with `TreeTagger` for French [53]. 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` [54]. 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 suppletive noun bases *myo N* (muscle)* and *carde N* (heart)*, which belong to the noun part-of-speech, and that it contains also the adjectival affix *-ique/ADJ*. At this step, the words are decomposed into their morphological components. We can observe that some bases can be decomposed further (*e.g. galactose* can be decomposed 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 obtained further to the decomposition are associated with (or "translated in") words from modern French. We use for this the resource presented in section IV-C (see 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)*

   We can observe that 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)*).

B. 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.

C. Extraction of layman paraphrases corresponding to technical terms

At this step, 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 French words are extracted as candidates for proposing the paraphrases. Additionally, the segments within sentences are also extracted as candidates for paraphrases. Currently, 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 maximal window size n is fixed to 10 words. Smaller or larger windows show lesser 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.*

For example, the sentence in (10) contains words *muscle (muscle)* and *coeur (heart)*, 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 bounded by these two words *muscle du coeur (heart muscle)*. The sentence and segment are expected to deliver the paraphrase of the technical term *myocardique (myocardial)*.

D. 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 a background in computational linguistics and morphology. We analyze the candidates for paraphrases from several points of view:

1) Are the French words corresponding to the components extracted correctly?
2) How easy are these paraphrases to be understood by laymen or by non-experts in medicine?
3) Do these French words provide valid candidates for paraphrases?

During the evaluation related to the third point (*Do these French words provide valid candidates for paraphrases?*), we distinguish four situations:

- the extraction is correct: *e.g. myocardique* paraphrased in *muscle du coeur (heart muscle)*;
- the extraction suffers from the incorrect morphological decomposition or from the wrong "translation" in French: *e.g. périanal (perianal)* is decomposed and "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 can collect several segments like *autour de 30 ans (around 30 years)*, which are not correct for our purpose;
- the extraction contains the correct paraphrase but should be post-processed. For instance, *spondylarthrose* is decomposed and "translated" in *vertèbre (vertebra)* and *arthrose (arthrosis)*. One of the paraphrases found is *arthrose que l'on ne voyait pas sur la vertèbre (arthrosis that was not seen on the vertebra)*. We think that this segment contains the necessary information, but must be transformed in *arthrose sur la vertèbre (arthrosis on the vertebra)* to be exploitable;
- the extraction is wrong and can provide no useful information.

This evaluation allows to estimate the precision of the results in three versions:

1) strict precision P_{strict}: only the correct extractions are considered;
2) weak precision P_{weak}: correct extractions together with extractions that are to be post-processed are considered;
3) very weak precision P_{vweak}: correct extractions, extractions that are to be post-processed and extractions with incorrect morphological processing are considered. The utility of this value is that it allows evaluating the percentage of the incorrect extractions.

As baseline, we use the definition contexts in which the neoclassical compounds occur. The definition are structures which contain two elements: *definiendum* (term to be defined) and *definiens* (the definition itself), like in *Myocardium is the muscular tissue of the heart*. Like paraphrases, the definitions provide explanations to terms, although in a more extensive form. We exploit the typical definition patterns [55] proposed in the literature, such as *is a*, *is defined as*. With this approach, we have to first map the technical term itself and then, if it occurs in a definition context, the whole sentence is extracted. Precision of the definitions is also evaluated manually.

VI. Results

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. In Table II, we present the results on extraction of the sentences and paraphrases from the corpora processed. For the three semantic types of terms (anatomy *ana.*, disorders *dis.*, and procedures *pro.*) from each corpus, we indicate the following information: number of different sentences extracted (*Number of sentences*), number of different terms (*Number of unique terms*).

In Table III, we indicate the evaluation of the extracted data: number of correct paraphrases (*correct*), number of paraphrases that are possibly correct (*pos. correct*), number of paraphrases which morphological analysis and "translation" should be improved (*morph. ana.*), and number of incorrect paraphrases (*incorrect*).

Corpus	*Number of sentences*	Number of unique terms
LesDiab		
ana.	15	7
dis.	71	30
pro.	10	5
DiabDoct		
ana.	721	35
dis.	2901	204
pro.	564	48
HT		
ana.	246	29
dis.	1233	133
pro.	678	42
Dos		
ana.	42	13
dis.	708	44
pro.	30	13

TABLE II. NUMBER OF EXTRACTIONS (SENTENCES AND UNIQUE TERMS).

Corpus	*Correct*	*Pos. correct*	*Morpho. ana.*	*Incorrect*
LesDiab				
ana.	3	3	1	8
dis.	32	7	2	30
pro.	4	-	6	-
DiabDoct				
ana.	227	40	100	354
dis.	1189	332	3	1
pro.	67	5	394	98
HT				
ana.	114	10	22	100
dis.	637	85	-	511
pro.	38	9	591	40
Dos				
ana.	12	3	2	25
dis.	466	98	1	135
pro.	13	2	12	3

TABLE III. RESULTS OF THE MANUAL EVALUATION OF THE EXTRACTED DATA.

In Table IV, we indicate the precision values: strict precision P_{strict}, weak precision P_{weak} and very weak precision P_{vweak}. They range from 6 to 100%.

Corpus	P_{strict}	P_{weak}	P_{vweak}
LesDiab			
ana.	20	40	47
dis.	45	55	58
pro.	40	40	100
DiabDoct			
ana.	32	37	51
dis.	40	52	53
pro.	12	13	83
HT			
ana.	46	50	59
dis.	52	59	59
pro.	6	7	59
Dos			
ana.	29	36	41
dis.	66	80	80
pro.	43	50	90

TABLE IV. PRECISION OBTAINED WITH DIFFERENT DATASETS.

VII. DISCUSSION

We develop the discussion along the following lines: the morphological analysis of terms (section VII-A), the processing of corpora (section VII-B), the extraction of paraphrases and their evaluation (section VII-C), the comparison with the baseline containing the definitions extracted (section VII-D), the comparison with the existing work (section VII-E), and analysis of non-analyzed terms (section VII-F).

A. Morphological analysis of terms

Among the words from the medical terminology, that have been analyzed morphologically, we can find compounds (*céphalgie (cephalalgia)*, *pharmacodépendance (pharmacodependency)* and words formed with affixes (*e.g. réadaptation* derived from *adaptation*, derived in its turn from *adapter*). Among the words, that have not been morphologically analyzed, we can find simple words (*e.g. abcès (abscess)*, *lèpre (leprosy)*, *cicatrice (scar)*) and words that contain bases and affixes currently not managed by `Dérif` (*e.g. pneumostrongylose (pneumostrongylosis)*, *lagophtalmie (lagophthalmos)*, *nécatorose (necatorosis)*). Among the decompositions generated by `Dérif`, we can find some cases with ambiguous analysis. Ambiguous decompositions occur when medical terms can be decomposed in several possible ways, among which only one is semantically correct. For instance, *posturographie (posturography)* is decomposed in: *[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 view point, this decomposition is very possible, although semantically it is weak. For this 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. This aspect of the method will be fixed through corrections in the resource.

B. Processing 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 also show very specific punctuation, which may be abusive, decorative, missing, etc. This seriously impedes the possibility to provide the correct segmentation into sentences, and means that, with the current method during the mapping of the decomposed terms with corpora, not sentences but bigger text segments may be considered. Within bigger segments, the semantic relations between the mapped components may be very weak or nonexistent, which may lead to incorrect extractions. We plan to combine the current method with the syntactic analysis in order to guarantee that stronger syntactic and semantic relations exist between the components.

C. Extraction of paraphrases and their evaluation

From the data presented in Tables II to IV, we can propose several observations:

- the *DiabDoct* corpus, that is the largest in our dataset, provides also the largest number of extractions (sentences, unique terms and paraphrases);
- among the three semantic types (anatomy, disorders and procedures), the number of paraphrases extracted for disorders is the largest in all corpora;
- the largest set of paraphrases, that suffer from incorrect morphological decomposition or "translation", is obtained for the procedure terms.

According to these observations, P_{strict} ranges between 20 to 46% for anatomy, 40 and 66% for disorders, and 6 to 43

for procedures. The P_{weak} values, that take into account the paraphrases that need post-processing, show the increase by 0 to 20%. The P_{vweak} values indicate that the anatomy terms show the largest rate (41 to 59%) of incorrect paraphrases. These are between 20 and 47 among the disorder terms, and between 0 to 41 between the procedure terms. We assume that the syntactic analysis may reduce the number of incorrect extractions and help to improve the current results.

Taken together, the proposed method allows to extract the paraphrases for 722 different terms from the corpora processed. These paraphrases are correct for 249 technical terms; while 299 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 paraphrases extracted in (11) to (16).

(11) *dorsalgie (dorsalgia)*: *douleur dans le dos (pain in the back)*

(12) *myélocyte (myelocyte)*: *cellules dans la moelle osseuse (cells of the bone marrow)*

(13) *lombalgie (lombalgia)*: *douleurs dans les reins (pain in kidney)*

(14) *gastralgie (gastralgia)*: *douleurs à l'estomac (stomach pain)*

(15) *desmorrhexie (desmorrhexia)*: *rupture des ligaments (ligamentous rupture)*

(16) *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 of 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)*.
- *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 tissue is absent, while necessary for its understanding. This situation, that brings the semantic underdetermination, has also been observed for other terms [50]. 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 more specific words, such as *increase* or *elevated* to extract correct paraphrases;
- *hétérotopie (heterotopia)* is "translated" in *autre (another)* and *endroit (place)*. The term means the displacement of an organ from its normal position. The morphological analysis only indicates 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 (*another place*) widely used in social media texts.

Among the incorrect extractions we can find more terms with non compositional semantics (such as *ostéodermie (osteoderm)*, *causalgie (causalgia)*, *adénoïde (adenoid)*, or *xanthochromie (xanthochromia)*) for which the extracted paraphrases can capture only part of the meaning, and 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)*). When the paraphrases are extracted from fora dedicated to a given medical topics, these paraphrases may remain concerned with this topics. This means that additional corpora must be used for improving the coverage of the paraphrases.

Taken together, we can consider that the currently proposed method allows extracting interesting candidates for the paraphrases of technical terms, that are indeed much easier to understand than the technical terms by themselves. 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 [56], [11]. These are inherent to the syntactic constructions extracted. The characteristics of these relations will be described and modeled in future work.

D. Comparison with the definition contexts

Corpus	*Nb of sentences*	*Nb of terms*	P_{strict}	P_{weak}
LesDiab				
ana.	1	1	0	100
dis.	7	7	43	57
pro.	1	1	0	0
DiabDoct				
ana.	27	16	7	7
dis.	264	92	17	22
pro.	36	16	11	25

TABLE V. EXTRACTION AND EVALUATION OF DEFINITIONS.

The definition patterns have been applied to two corpora: *LesDiab* and *DiabDoct*. In Table V, we indicate the total number of definitions extracted and the number of unique terms. We then indicate the precision values: strict and weak. Although in a different perspective, the strong precision consider only correct definitions, while the weak precision also accepts possibly correct definitions. We explain these two precision types later in this section.

We extract 28 definition contexts for the anatomy terms, 37 definition contexts for the procedure terms, and 271 definition contexts for the pathology terms. The pattern `est un` (`is a`) is the most frequently recognized. It detects a great number of definition contexts, which may provide with the definitions of the terms, but also with the points of view

given by the forum users. We present and discuss some of these contexts below. Other patterns, *i.e.* `également appelé` (`also called`) and `peut être défini comme` (`can be defined as`), are also recognized but with a lesser frequency. Precision of the `est un` (`is a`) pattern is weak but it provides a large number of propositions, while other patterns, when they are matched, show a greater precision.

The contexts deemed as relevant provide definitions for 38 different terms:

1) neoclassical compounds: *acidocétose (ketoacidosis)*, *hypoglycémie (hypoglycemia)*, *angiographie (angiography)*, *hypokaliémie (hypokaliemia)*,
2) affixed terms: *curetage (scraping)*, *capsulite (capsulitis)*, *arthrose (arthrosis)*, *glaucome (glaucoma)*, *durillon (callus)*, *pré-diabète (prediabetes)*,
3) and morphologically non constructed terms: *cataracte (cataract)*, *imptigo (impetigo)*, *zona (shingles)*.

In relation to the main method exploiting the structure of neoclassical compounds, only the first type of terms (*i.e.* neoclassical compounds) is directly comparable.

The definition contexts may correspond to intentional (analytical) definitions:

- *L'hypoglycémie est un manque de sucre dans l'organisme (Hypoglycemia is a lack of sugar in the organism)*
- *Une septicémie est un empoisonnement du sang du à un microbe (Septicemia is a blood poisoning due to a microbe)*
- *Le curetage est un nettoyage en profondeur d'une gencive inflammée (Scraping is an in depth cleaning of an infected gum)*

but also to extensional (examples, signs...) definitions:

- *Pour un être humain adulte, une hypoglyémie est une glycémie inférieure à 0,8 g/L (For an adult human, the hypoglycemia is a glycemia inferior to 0,8 g/L)*
- *Les signes classiques annonciateurs de l'hypoglycmie sont des sueurs, pleur, palpitations, fringales en particulier (Typical signs warning about the hypoglycemia are sweating, pallor, palpitations, and raging hunger especially)*

or to their combination:

- *L'impétigo est une infection cutanée, qui provoque des pustules qui dégénèrent en croûtes jaunâtres, l'impétigo est due à... (Impetigo is a skin infection, which causes pustules which degenerate into yellow scab, impetigo is due to...)*
- *L'angiographie est un examen des yeux permettant de détecter des microanévrismes (Angiography is an examination of eyes which allows to detect microanevrisms)*
- *Or, une baisse de potassium dans le sang, appelée hypokaliémie, peut se traduire par... (And yet, the decrease of potassium in blood, called hypokaliemia, can result in...)*

Among the contexts deemed as possibly relevant, we can find what we call the *points of view* definitions, such as:

- *La mélancolie est une douceur qui nous berce (Melancholia is a gentleness that lulls us)*
- *Une injection est une agression, qui sauve, mais c'est quand même une agression (Injection is an aggression, which saves, but an aggression nonetheless)*

Such definitions give very personal perception of the terms.

By comparison with the main method, we can easily observe that paraphrases cover a larger amount of terms and sentences (Table II). This fact is expected because with the paraphrase method, we do not rely on the occurrence of the term to be paraphrased: we just need its components to occur in the text. Concerning precision, its values are also better with the main method (Table IV). As for the usability of the definitions, the intentional definitions suit the best the purpose of the current study. Still, they need to be post-processed and to be transformed from definitions to paraphrases. In a different perspective, both intentional and extensional definitions can be used for explaining the medical notions. In this case, they do not build a lexicon like in examples in (4), but can be used nevertheless for providing natural language definitions of terms.

E. Comparison with the existing work

We can compare the obtained results with the results presented in three previous experiments ([39], [40], [38]):

- concerning the number of the extracted paraphrases:
  - in our work, we extract 722 paraphrases, among which 299 are correct,
  - in [39], 65 and 82 paraphrases are extracted,
  - in [40], 109 paraphrases are extracted,
  - in [38], 152 paraphrases are extracted;
- concerning the precision values:
  - in our work, according to types of terms, we obtain precision between 20 and 59% for the anatomy terms, between 40 and 80% for the disorder terms, and between 6 and 100% for the procedure terms,
  - in [39], precision is 67% and 60%,
  - in [40], precision is 66%,
  - in [38], precision is 58%.

Notice that in the existing studies, only in one of them [38] is based on the exploitation of terms from an existing medical terminology. The two other studies exploit the content of the corpora and no link is done with the existing terminology. By comparison with this work [38], we provide a better coverage of the paraphrased terms.

Still, it remains difficult to compute the recall values. One possibility is to consider the whole set of the available terms (4,170 terms), but all these terms cannot be analyzed morphologically and, for this reason, they cannot be processed by the proposed method. Another possibility for computing the recall is to consider the terms that can be analyzed morphologically (3,030 terms). In this case, the recall value is close to 10% with the correct paraphrases (299 terms),

while it is close 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. Yet another possibility is to consider that recall is to be computed within the paraphrases extracted (722 terms), especially because our method is permissive and should not miss the paraphrases. In this case, recall for the correct unique terms paraphrased is 41%. When the evaluation is done at the level of the paraphrase occurrences, the recall values are: 40% for the anatomy terms, 58% for the disorder terms, and 11% for the procedure terms. The average recall is 47% then. Finally, the last possibility to compute recall is to annotate the corpus with the paraphrases of compounds. This evaluation is not done in the current work, but will be done in the future. The main difficulty with this kind of evaluation is to manage the terms to be paraphrased and the correspondences between Latin and Greek bases and their aquivalences in the modern language, such as in {*card*, *heart*}, {*myo*, *muscle*}, {*ectomie*, *removal*}.

F. Analysis of non-analyzed terms

If the coverage of the terms that have been associated with paraphrases is quite important (722 terms), several other terms do not receive the paraphrases. Some of them are presented in (17). One reason may be that some terms, like *hémidesmosome* or *hémohistioblaste*, contain several (more than two) components, which may make it more difficult to map them to the text and to extract the paraphrases. Other terms may contain prefixes or words that naturally occur less frequently. Nevertheless, we expect that using additional and larger corpus will provide additional paraphrases and complete the paraphrases extracted.

(17) *leptoméningé (leptomeninge)*: *affaibli (impaired)*, *méningé (meningeal)*,
hémipénis (hemipenis): *pénis (penis)*, *demi (half)*,
otolithe (otolith): (*calcul (stone)*, *oreille (ear)*,
hémidesmosome (hemidesmosome): *corpuscule (corpuscle)*, *demi (half)*, *ligament (ligament)*,
hémohistioblaste (hemohistioblast): *cellule embryonnaire (stem cells)*, *tissu (tissue)*, *sang (blood)*

VIII. Conclusions and Future Work

We proposed 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 with the French data. 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 paraphrases for up to 299 technical compound 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 correctly 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 the existing automatic approaches, as they do not present clear formal similarity with their paraphrases.

One of the current difficulties 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. As we indicated in examples in (8), some words are not fully decomposed by the `DériF` analyzer, which should be improved in future work and would allow to improve the coverage of the extractions. In future work, we will also take into account syntactically complex terms and not only simple words. We assume that the proposed method can be applied to other languages provided that the morphological decomposition or analysis can be performed. The very objective of our work is to exploit the resource created for the simplification of medical texts.

Acknowledgment

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

[1] AMA, "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*, vol. 281, no. 6, pp. 552–7, 1999.

[2] A. McCray, "Promoting health literacy," *J of Am Med Infor Ass*, vol. 12, pp. 152–163, 2005.

[3] G. Eysenbach, "Poverty, human development, and the role of ehealth," *J Med Internet Res*, vol. 9, no. 4, p. e34, 2007.

[4] Oregon Evidence-based Practice Center, "Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved," Agency for healthcare research and quality, Tech. Rep., 2008.

[5] V. Patel, T. Branch, and J. Arocha, "Errors in interpreting quantities as procedures : The case of pharmaceutical labels," *International journal of medical informatics*, vol. 65, no. 3, pp. 193–211, 2002.

[6] M. Williams, R. Parker, D. Baker, N. Parikh, K. Pitkin, W. Coates, and J. Nurss, "Inadequate functional health literacy among patients at two public hospitals," *JAMA*, vol. 274, no. 21, pp. 1677–82, 1995.

[7] G. Berland, M. Elliott, L. Morales, J. Algazy, R. Kravitz, M. Broder, D. Kanouse, J. Munoz, J. Puyol, M. Lara, K. Watkins, H. Yang, and E. McGlynn, "Health information on the internet. accessibility, quality, and readability in english ans spanish," *JAMA*, vol. 285, no. 20, pp. 2612–2621, 2001.

[8] D. Hargrave, U. Bartels, L. Lau, C. Esquembre, and E. Bouffet, "évaluation de la qualité de l'information médicale francophone accessible au public sur internet : application aux tumeurs cérébrales de l'enfant," *Bulletin du Cancer*, vol. 90, no. 7, pp. 650–5, 2003.

[9] S. Kusec, "Les sites web relatifs au diabète, sont-ils lisibles ?" *Dibète et société*, vol. 49, no. 3, pp. 46–48, 2004.

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

[11] G. Booij, *Construction Morphology*. Oxford: Oxford University Press, 2010.

[12] C. Iacobini, "Distinguishing derivational prefixes from initial combining forms," in *First mediterranean conference of morphology*, Mytilene, Island of Lesbos, Greece, septembre 1997.

[13] D. Amiot and G. Dal, "Integrating combining forms into a lexeme-based morphology," in *Mediterranean Morphology Meeting (MMM5)*, 2005, pp. 323–336.

[14] X. Zeng and B. Parmanto, "Evaluation of web accessibility of consumer health information websites," in *AMIA 2003*, 2003, pp. 743–7.

[15] T. François, "Les apports du traitements automatique du langage la lisibilit du franais langue trangre," PhD thesis, Universit Catholique de Louvain, Louvain, 2011.

[16] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 23, pp. 221–233, 1948.

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

[18] W. H. Dubay, "The principles of readability," *Impact Information*, 2004, available at *http://almacenplantillasweb.es/wp-content/uploads/2009/11/The-Principles-of-Readability.pdf*.

[19] R. Rudd, B. Moeykens, and T. Colton, *Annual Review of Adult Learning and Literacy*, 1999, p. ch 5.

[20] E. Rudd, "Needed action in health literacy," *J Health Psychol*, vol. 18, no. 8, pp. 1004–10, 2013.

[21] D. Kokkinakis and M. Toporowska Gronostaj, "Comparing lay and professional language in cardiovascular disorders corpora," in *WSEAS Transactions on BIOLOGY and BIOMEDICINE*, A. Pham T., James Cook University, Ed., 2006, pp. 429–437.

[22] M. Poprat, K. Markó, and U. Hahn, "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*, Maastricht, 2006, pp. 503–508.

[23] L. Goeuriot, N. Grabar, and B. Daille, "Caractérisation des discours scientifique et vulgarisé en français, japonais et russe," in *TALN*, 2007, pp. 93–102.

[24] T. Miller, G. Leroy, S. Chatterjee, J. Fan, and B. Thoms, "A classifier to evaluate language specificity of medical documents." in *HICSS*, 2007, pp. 134–140.

[25] J. Chmielik and N. Grabar, "Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques," *TAL*, vol. 51, no. 2, pp. 151–179, 2011.

[26] Y. Wang, "Automatic recognition of text difficulty from consumers health information," in *Computer-Based Medical Systems*, IEEE, Ed., 2006, pp. 131–136.

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

[28] G. Leroy, S. Helmreich, J. Cowie, T. Miller, and W. Zheng, "Evaluating online health information: Beyond readability formulas," in *AMIA 2008*, 2008, pp. 394–8.

[29] T. François and C. Fairon, "Les apports du TAL à la lisibilité du français langue étrangère," *TAL*, vol. 54, no. 1, pp. 171–202, 2013.

[30] L. Specia, S. Jauhar, and R. Mihalcea, "Semeval-2012 task 1: English lexical simplification," in **SEM 2012*, 2012, pp. 347–355.

[31] R. Sinha, "Unt-simprank: Systems for lexical simplification ranking," in **SEM 2012*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 493–496. [Online]. Available: http://www.aclweb.org/anthology/S12-1069

[32] S. Jauhar and L. Specia, "Uow-shef: Simplex – lexical simplicity ranking based on contextual and psycholinguistic features," in **SEM 2012*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 477–481. [Online]. Available: http://www.aclweb.org/anthology/S12-1066

[33] A. Johannsen, H. Martínez, S. Klerke, and A. Søgaard, "Emnlp@cph: Is frequency all there is to simplicity?" in **SEM 2012*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 408–412. [Online]. Available: http://www.aclweb.org/anthology/S12-1054

[34] A. Ligozat, C. Grouin, A. Garcia-Fernandez, and D. Bernhard, "Annlor: A naïve notation-system for lexical outputs ranking," in **SEM 2012*, 2012, pp. 487–492.

[35] M. Amoia and M. Romanelli, "Sb: mmsystem - using decompositional semantics for lexical simplification," in **SEM 2012*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 482–486. [Online]. Available: http://www.aclweb.org/anthology/S12-1067

[36] Q. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *JAMIA*, vol. 13, pp. 24–29, 2006.

[37] D. Lindberg, B. Humphreys, and A. McCray, "The unified medical language system," *Methods Inf Med*, vol. 32, no. 4, pp. 281–291, 1993.

[38] N. Elhadad and K. Sutaria, "Mining a lexicon of technical terms and lay equivalents," in *BioNLP*, 2007, pp. 49–56.

[39] L. Deléger and P. Zweigenbaum, "Paraphrase acquisition from comparable medical corpora of specialized and lay texts," in *AMIA 2008*, 2008, pp. 146–50.

[40] B. Cartoni and L. Deléger, "Dcouverte de patrons paraphrastiques en corpus comparable: une approche base sur les n-grammes," in *TALN*, 2011.

[41] R. Messai, Q. Zeng, M. Mousseau, and M. Simonet, "Building a bilingual french-english patient-oriented terminology for breast cancer," in *MedNet*, 2006.

[42] U. Hahn, M. Honeck, M. Piotrowsky, and S. Schulz, "Subword segmentation - leveling out morphological variations for medical document retrieval," in *AMIA*, 229-33, 2001.

[43] S. Fernández-Silva, J. Freixa, and M. Cabré, "A proposed method for analysing the dynamics of cognition through term variation," *Terminology*, vol. 17, no. 1, pp. 49–73, 2011.

[44] A. Max, H. Bouamor, and A. Vilnat, "Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs," in *EMNLP*, 2012, pp. 721–31.

[45] C. Lovis, P.-A. Michel, R. Baud, and J.-R. Scherrer, "Word segmentation processing: a way to exponentially extend medical dictionaries," in *Medical Informatics in Europe (MIE)*, 1995, pp. 28–32.

[46] S. Schulz, M. Romacker, P. Franz, A. Zaiss, R. Klar, and U. Hahn, "Towards a multilingual morpheme thesaurus for medical free-text retrieval," in *Medical Informatics in Europe (MIE)*, 1999, pp. 891–4.

[47] M. G. Pacak, L. M. Norton, and G. S. Dunham, "Morphosemantic analysis of -itis forms in medical language," *Methods in Medical Informatics (MIM)*, vol. 19, no. 2, pp. 99–105, 1980.

[48] P. Dujols, P. Aubas, C. Baylon, and F. Grémy, "Morphosemantic analysis and translation of medical compound terms," *Methods in Informatics and Medicin (MIM)*, vol. 30, pp. 30–35, 1991.

[49] S. Wolff, "Automatic coding of medical vocabulary," in *Medical Language Processing. Computer Management of Narrative Data*, N. Sager, C. Friedman, and M. S. Lyman, Eds. New-York: Addison-Wesley, 1987, ch. 7, pp. 145–162.

[50] A. T. McCray, A. C. Browne, and D. Moore, "The semantic structure of neo-classical compounds," in *Proceedings of the Annual SCAMC*, 1988, pp. 165–168.

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

[52] P. Zweigenbaum and N. Grabar, "Corpus-based associations provide additional morphological variants to medical terminologies," in *AMIA*, 2003.

[53] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *ICNMLP*, Manchester, UK, 1994, pp. 44–49.

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

[55] M. Péry-Woodley and J. Rebeyrolle, "Domain and genre in sublanguage text: definitional microtexts in three corpora," in *First International Conference on Language Resources and Evaluation*, 1998, pp. 987–992.

[56] F. Namer and P. Zweigenbaum, "Acquiring meaning for French medical terminology: contribution of morphosemantics," in *Annual Symposium of the American Medical Informatics Association (AMIA)*, San-Francisco, 2004.