Automatic diagnosis of understanding of medical words

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

Dany Amiot
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq, France
dany.amiot@univ-lille3.fr

Abstract

Within the medical field, very specialized terms are commonly used, while their understanding by laymen is not always successful. We propose to study the understandability of medical words by laymen. Three annotators are involved in the creation of the reference data used for training and testing. The features of the words may be linguistic (*i.e.*, number of characters, syllables, number of morphological bases and affixes) and extra-linguistic (*i.e.*, their presence in a reference lexicon, frequency on a search engine). The automatic categorization results show between 0.806 and 0.947 F-measure values. It appears that several features and their combinations are relevant for the analysis of understandability (*i.e.*, syntactic categories, presence in reference lexica, frequency on the general search engine, final substring).

1 Introduction

The medical field has deeply penetrated our daily life, which may be due to personal or family health condition, watching TV and radio broadcasts, reading novels and journals. Nevertheless, the availability of this kind of information does not guarantee its correct understanding, especially by laymen, such as patients. The medical field has indeed a specific terminology (*e.g.*, *abdominoplasty*, *hepatic*, *dermabrasion* or *hepatoduodenostomy*) commonly used by medical professionals. This fact has been highlighted in several studies dedicated for instance to the understanding of pharmaceutical labels (Patel et al., 2002), of information provided by websites (Rudd et al., 1999; Berland et al., 2001; McCray, 2005; Oregon Evidence-based Practice Center, 2008), and more generally the understanding between patients and medical doctors (AMA, 1999; McCray, 2005; Jucks and Bromme, 2007; Tran et al., 2009).

We propose to study the understanding of words used in the medical field, which is the first step towards the simplification of texts. Indeed, before the simplification can be performed, it is necessary to know which textual units may show understanding difficulty and should be simplified. We work with data in French, such as provided by an existing medical terminology. In the remainder, we present first some related work, especially from specialized fields (section 2). We then introduce the linguistic data (section 4) and methodology (section 5) we propose to test. We present and discuss the results (section 6), and conclude with some directions for future work (section 7).

2 Studying the understanding of words

The understanding (of words) may be seen as a scale going from *I can understand* to *I cannot understand*, and containing one or more intermediate positions (*i.e.*, *I am not sure*, *I have seen it before but do not remember the meaning*, *I do not know but can interpret*). Notice that it is also related to the ability to provide correct explanation and use of words. As we explain later, we consider words out of context and use a three-position scale. More generally, understanding is a complex notion closely linked to several other notions studied in different research fields. For instance, lexical complexity is studied in linguistics and gives clues on lexical processes involved, that may impact the word understanding (section 2.1). Work in psycholinguistics is often oriented on study of word opacity and the mental processes involved in their understanding (Jarema et al., 1999; Libben et al., 2003). Readability provides a set of methods to compute and quantify the understandability of words (section 2.3). The specificity of words to specialized areas is another way to capture their understandability (section 2.2). Finally, lexical

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 11–20,
Gothenburg, Sweden, April 26-30 2014.

simplification aims at providing simpler words to be used in a given context (section 2.3).

2.1 Linguistics

In linguistics, the question is closely related to lexical complexity and compoundings. It has been indeed observed that at least five factors, linguistic and extra-linguistic, may be involved in the semantic complexity of the compounds. One factor is related to the knowledge of the components of the complex words. Formal (how the words, such as *aérenchyme*, can be segmented) and semantic (how the words can be understood and used) points of view can be distinguished. A second factor is that complexity is also due to the variety of morphological patterns and relations among the components. For instance, *érythrocyte* (*erythrocyte*) and *ovocyte* (*ovocyte*) instantiate the *[N1N2]* pattern in which *N2* (*cyte*) can be seen as a constant element (Booij, 2010), although the relations between *N1* and *N2* are not of the same type in these two compounds: in *érythrocyte*, *N1 érythr(o)* denotes a property of *N2* (color), while in *ovocyte*, *N1 ovo* (*egg*) corresponds to a specific development stage of female cells. Another factor appears when some components are polysemous, within a given field (*i.e.*, medical field) or across the fields. For instance, *aér(o)* does not always convey the same meaning: in *aérocèle*, *aér-* denotes 'air' (*tumefaction (cèle) formed by an air infiltration*), but not in *aérasthénie*, which refers to an *asthenia (psychic disorder)* observable among jet pilots. Yet another factor may be due to the difference in the order of components: according to whether the compounding is standard (in French, the main semantic element is then on the left, such as in *pneu neige* (*snow tyre*), which is fundamentally a *pneu* (*tyre*)) or neoclassical (in French, the main semantic element is then on the right, such as *érythrocyte*, which is a kind of cyte *cell / corpuscle* with red color). It is indeed complicated for a user without medical training to correctly interpret a word that he does not know and for which he cannot reuse the existing standard compounding patterns. This difficulty is common to all Roman languages (Iacobini, 2003), but not to Germanic languages (Lüdeling et al., 2002). Closely related is the fact that with neoclassical compounds, a given component may change its place according to the global semantics of the compounds, such as *path-* in *pathology, polyneuropathe, cardiopathy*. Finally, the formal similarity between some derivation processes (such as the derivation in *-oide*, like in *lipoid*) and neoclassical compounding (such as *-ase* in *lipase*), which apply completely different interpretation patterns (Iacobini, 1997; Amiot and Dal, 2005), can also make the understanding more difficult.

2.2 Terminology

In the terminology field, the automatic identification of difficulty of terms and words remains implicit, while this notion is fundamental in terminology (Wüster, 1981; Cabré and Estopà, 2002; Cabré, 2000). The specificity of terms to a given field is usually studied. The notion of understandability can be derived from it. Such studies can be used for filtering the terms extracted from specialized corpora (Korkontzelos et al., 2008). The features exploited include for instance the presence and the specificity of pivot words (Drouin and Langlais, 2006), the neighborhood of the term in corpus or the diversity of its components computed with statistical measures such as C-Value or PageRank (Daille, 1995; Frantzi et al., 1997; Maynard and Ananiadou, 2000). Another possibility is to check whether lexical units occur within reference terminologies and, if they do, they are considered to convey specialized meaning (Elhadad and Sutaria, 2007).

2.3 NLP studies

The application of the readability measures is another way to evaluate the complexity of words and terms. Among these measures, it is possible to distinguish classical readability measures and computational readability measures (François, 2011). Classical measures usually rely on number of letters and/or of syllables a word contains and on linear regression models (Flesch, 1948; Gunning, 1973), while computational readability measures may involve vector models and a great variability of features, among which the following have been used to process the biomedical documents and words: combination of classical readability formulas with medical terminologies (Kokkinakis and Toporowska Gronostaj, 2006); n-grams of characters (Poprat et al., 2006), manually (Zheng et al., 2002) or automatically (Borst et al., 2008) defined weight of terms, stylistic (Grabar et al., 2007) or discursive (Goeuriot et al., 2007) features, lexicon (Miller et al., 2007), morphological features (Chmielik and Grabar, 2011), combi-

Categories	*A1 (%)*	*A2 (%)*	*A3 (%)*	*Unanimity (%)*	*Majority (%)*
1. I can understand	8,099 (28)	8,625 (29)	7,529 (25)	5,960 (26)	7,655 (27)
2. I am not sure	1,895 (6)	1,062 (4)	1,431 (5)	61 (0.3)	597 (2)
3. I cannot understand	19,647 (66)	19,954 (67)	20,681 (70)	16,904 (73.7)	20,511 (71)
Total annotations	29,641	29,641	29,641	22,925	28,763

Table 1: Number (and percentage) of words assigned to reference categories by three annotators (A1, A2 and A3), and in the derived datasets *unanimity* and *majority*.

nations of different features (Wang, 2006; Zeng-Treiler et al., 2007; Leroy et al., 2008).

Specific task has been dedicated to the lexical simplification within the *SemEval* challenge in 2012[1]. 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). The participants applied rule-based and/or machine learning systems. Combinations of various features have been used: lexicon from spoken corpus and Wikipedia, Google n-grams, WordNet (Sinha, 2012); word length, number of syllables, latent semantic analysis, mutual information and word frequency (Jauhar and Specia, 2012); Wikipedia frequency, word length, n-grams of characters and of words, random indexing and syntactic complexity of documents (Johannsen et al., 2012); n-grams and frequency from Wikipedia, Google n-grams (Ligozat et al., 2012); WordNet and word frequency (Amoia and Romanelli, 2012).

3 Aims of the present study

We propose to investigate how the understandability of French medical words can be diagnosed with NLP methods. We rely on the reference annotations performed by French speakers without medical training, which we associate with patients. The experiments performed rely on machine learning algorithms and a set of 24 features. The medical words studied are provided by an existing medical terminology.

4 Linguistic data and their preparation

The linguistic data are obtained from the medical terminology Snomed International (Côté, 1996). This terminology's aim is to describe the whole medical field. It contains 151,104 medical terms structured into eleven semantic axes such as disorders and abnormalities, procedures, chemical products, living organisms, anatomy, social status, etc. We keep here five axes related to the main medical notions (disorders, abnormalities, procedures, functions, anatomy). The objective is not to consider axes such as chemical products (*trisulfure d'hydrogène (*hydrogen sulfide*)*) and living organisms (*Sapromyces, Acholeplasma laidlawii*) that group very specific terms hardly known by laymen. The 104,649 selected terms are tokenized and segmented into words (or tokens) to obtain 29,641 unique words: *trisulfure d'hydrogène* gives three words (*trisulfure, de, hydrogène*). This dataset contains compounds (*abdominoplastie (*abdominoplasty*), dermabrasion (*dermabrasion*)*), constructed (*cardiaque (*cardiac*), acineux (*acinic*), lipoïde (*lipoid*)*) and simple (*acné (*acne*), fragment (*fragment*)*) words. These data are annotated by three speakers 25-40 year-old, without medical training, but with linguistic background. We expect the annotators to represent the average knowledge of medical words amongst the population as a whole. The annotators are presented with a list of terms and asked to assign each word to one of the three categories: (1) I can understand the word; (2) I am not sure about the meaning of the word; (3) I cannot understand the word. The assumption is that the words, which are not understandable by the annotators, are also difficult to understand by patients. These manual annotations correspond to the reference data (Table 1).

5 Methodology

The proposed method has two aspects: generation of the features associated to the analyzed words and a machine learning system. The main research question is whether the NLP methods can distinguish between understandable and non-understandable medical words and whether they can diagnose these two categories.

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

5.1 Generation of the features

We exploit 24 linguistic and extra-linguistic features related to general and specialized languages. The features are computed automatically, and can be grouped into ten classes:

Syntactic categories. Syntactic categories and lemmas are computed by TreeTagger (Schmid, 1994) and then checked by Flemm (Namer, 2000). The syntactic categories are assigned to words within the context of their terms. If a given word receives more than one category, the most frequent one is kept as feature. Among the main categories we find for instance nouns, adjectives, proper names, verbs and abbreviations.

Presence of words in reference lexica. We exploit two reference lexica of the French language: TLFi[2] and *lexique.org*[3]. TLFi is a dictionary of the French language covering XIX and XX centuries. It contains almost 100,000 entries. *lexique.org* is a lexicon created for psycholinguistic experiments. It contains over 135,000 entries, among which inflectional forms of verbs, adjectives and nouns. It contains almost 35,000 lemmas.

Frequency of words through a non specialized search engine. For each word, we query the Google search engine in order to know its frequency attested on the web.

Frequency of words in the medical terminology. We also compute the frequency of words in the medical terminology Snomed International.

Number and types of semantic categories associated to words. We exploit the information on the semantic categories of Snomed International.

Length of words in number of their characters and syllables. For each word, we compute the number of its characters and syllables.

Number of bases and affixes. Each lemma is analyzed by the morphological analyzer Dérif (Namer and Zweigenbaum, 2004), adapted to the treatment of medical words. It performs the decomposition of lemmas into bases and affixes known in its database and it provides also semantic explanation of the analyzed lexemes. We exploit the morphological decomposition information (number of affixes and bases).

Initial and final substrings of the words. We compute the initial and final substrings of different length, from three to five characters.

Number and percentage of consonants, vowels and other characters. We compute the number and the percentage of consonants, vowels and other characters (*i.e.*, hyphen, apostrophe, comas).

Classical readability scores. We apply two classical readability measures: Flesch (Flesch, 1948) and its variant Flesch-Kincaid (Kincaid et al., 1975). Such measures are typically used for evaluating the difficulty level of a text. They exploit surface characteristics of words (number of characters and/or syllables) and normalize these values with specifically designed coefficients.

5.2 Machine learning system

The machine learning algorithms are used to study whether they can distinguish between words understandable and non-understandable by laymen and to study the importance of various features for the task. The functioning of machine learning algorithms is based on a set of positive and negative examples of the data to be processed, which have to be described with suitable features such as those presented above. The algorithms can then detect the regularities within the training dataset to generate a model, and apply the generated model to process new unseen data. We apply various algorithms available within the WEKA (Witten and Frank, 2005) platform.

The annotations provided by the three annotators constitute our reference data. We use on the whole five reference datasets (Table 1): 3 sets of separate annotations provided by the three annotators (29,641 words each); 1 *unanimity* set, on which all the annotators agree (n=22,925); 1 *majority* set, for which we can compute the majority agreement (n=28,763). By definition, the two last datasets should present a better coherence and less annotation ambiguity because some ambiguities have been resolved by unanimity or by majority vote.

5.3 Evaluation

The inter-annotator agreement is computed with the Cohen's Kappa (Cohen, 1960), applied to pairs of annotators, which values are then leveraged to obtain the unique average value; and Fleiss' Kappa (Fleiss and Cohen, 1973), suitable for processing data provided by more than two annotators. The interpretation of the scores are for instance (Landis and Koch, 1977): substantial agreement between 0.61 and 0.80, almost perfect agreement between 0.81 and 1.00.

[2] http://www.atilf.fr/

[3] http://www.lexique.org/

With machine learning, we perform a ten-fold cross-validation, which means that the evaluation test is performed ten times on different randomly generated test sets (1/10 of the whole dataset), while the remaining 9/10 of the whole dataset is used for training the algorithm and creating the model. In this way, each word is used during the test step. The success of the applied algorithms is evaluated with three classical measures: $\mathcal{R}$ recall, $\mathcal{P}$ precision and $\mathcal{F}$ F-measure. In the perspective of our work, these measures allow evaluating the suitability of the methodology to the distinction between understandable and non-understandable words and the relevance of the chosen features.

The baseline corresponds to the assignment of words to the biggest category, *e.g.*, *I cannot understand*, which represents 66 to 74%, according to datasets. We can also compute the gain, which is the effective improvement of performance P given the baseline BL (Rittman, 2008): $\frac{P-BL}{1-BL}$.

6 Automatic analysis of understandability of medical words: Results and Discussion

We address the following aspects: annotations (inter-annotator agreement, assignment of words to three categories), quantitative results provided by the machine learning algorithms, impact of the individual features on the distinction between categories, and usefulness of the method.

6.1 Annotations and inter-annotator agreement

The time needed for performing the manual reference annotations depends on annotators and ranges from 3 to 6 weeks. The annotation results presented in Table 1 indicate that the annotators 1 and 2 often provide similar results on their understanding of the medical words, while for the third annotator the task appears to be more difficult as he indicates globally a higher number of non-understandable words. The non-understandable words are the most frequent for all annotators and cover 66 to 70% of the whole dataset. The inter-annotator agreement shows substantial agreement: Fleiss' Kappa 0.735 and Cohen's Kappa 0.736. This is a very good result, especially when working with linguistic data for which the agreement is usually difficult to obtain.

The evolution of annotations per category (Figure 1), such as provided by the annotators, can distinguish easily between the three categories: (1) the most frequently chosen category is *I cannot understand* and it grows rapidly with new words; (2) the next most frequently chosen category is *I can understand*, although it grows more slowly; (3) the third category, which gathers the words on which the annotators show some hesitation, is very small. Given the proximity between the lines in each category, we can conclude that the annotators have similar difficulties in understanding the words from the dataset.

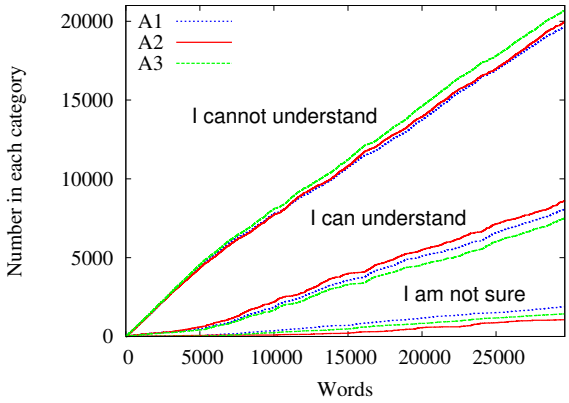

Figure 1: Evolution of the annotations within the reference data.

6.2 Quantitative results obtained with machine learning

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
J48	0.876	0.889	0.881
RandomForest	0.880	0.892	0.884
REPTree	0.874	0.890	0.879
DecisionTable	0.872	0.891	0.880
LMT	0.876	0.895	0.884
SMO	0.858	0.876	0.867

Table 2: Performance obtained on the *majority* dataset with various algorithms.

We tested several machine learning algorithms to discover which of them are the most suitable to the task at hand. In Table 2, with results computed on the *majority* dataset, we can observe that the algorithms provide with similar performance (between 0.85 and 0.90 $\mathcal{P}$ and $\mathcal{R}$). In the remaining of the paper, we present results obtained with J48 (Quinlan, 1993). Table 3 shows $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{F}$ values for the five datasets: three annotators, majority and unanimity datasets. We can observe

that, among the three annotators, it is easier to reproduce the annotations of the third annotator: we gain then 0.040 with $\mathcal{F}$ comparing to the two other annotators. The results become even better with the majority dataset ($\mathcal{F}$=0.881), and reach $\mathcal{F}$ up to 0.947 on the unanimity dataset. As we expected, these two last datasets present less annotation ambiguity. The best categorization results are observed with *I can understand* and *I cannot understand* categories, while the *I am not sure* category is poorly managed by machine learning algorithms. Because this category is very small, the average performance obtained on all three categories remains high.

	A1	*A2*	*A3*	*Una.*	*Maj.*
$\mathcal{P}$	0.794	0.809	0.834	0.946	0.876
$\mathcal{R}$	0.825	0.826	0.862	0.949	0.889
$\mathcal{F}$	0.806	0.814	0.845	0.947	0.881

Table 3: J48 performance obtained on five datasets (A1, A2, A3, *unanimity* and *majority*).

In Table 4, we indicate the gain obtained by J48 compared to baseline: it ranges from 0.13 to 0.20, which is a good improvement, despite the category *I am not sure* that is difficult to discriminate. We also indicate the accuracy obtained on these datasets.

	A1	*A2*	*A3*	*Una.*	*Maj.*
BL	0.66	0.67	0.70	0.74	0.71
$\mathcal{F}$	0.806	0.814	0.845	0.947	0.881
gain	0.14	0.13	0.14	0.20	0.16
Acc.	0.825	0.826	0.862	0.948	0.889

Table 4: Gain obtained for $\mathcal{F}$ by J48 on five datasets (A1, A2, A3, *unanimity* and *majority*).

6.3 Impact of individual features on understandability of medical words

To observe the impact of individual features, we did several iterations of experiments during which we incrementally increased the set of features: we started with one feature and then, at each iteration, we added one new feature, up to the 24 features available. We tried several random orders. The test presented here is done again on the *majority* dataset. Figures 2 present the results obtained in terms of $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{F}$. Globally, we can observe that some features show positive impact while others show negative or null impact:

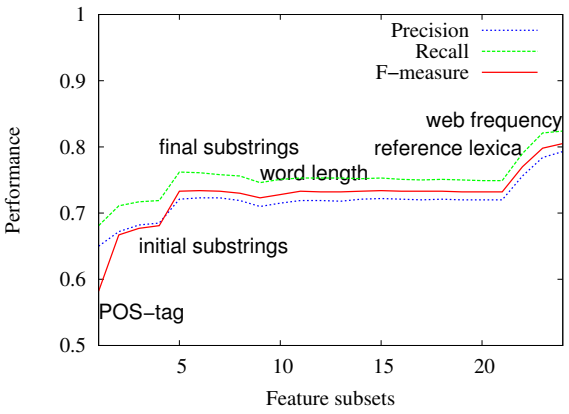

Figure 2: Impact of individual features.

- with the syntactic categories (POS-tags) alone we obtain $\mathcal{P}$ and $\mathcal{R}$ between 0.65 and 0.7. The performance is then close to the baseline performance. Often, proper names and abbreviations are associated with the non-understandable words. There is no difference between TreeTagger alone and the combination of TreeTagger with Flemm;

- the initial and final substrings have positive impact. Among the final substrings, those with three and four characters (ie, *-omie* of *-tomie* (*meaning* cut), *-phie* of *-rraphie* (*meaning* stitch), *-émie* (*meaning* blood)) show positive impact, but substrings with five characters have negative impact and the previously gained improvement is lost. We may conclude that the five-character long final substrings may be too specific;

- the length of words in characters have negative impact on the categorization results. There seems to be no strong link between this feature and the understanding of words: short and long words may be experienced as both understandable or not by annotators;

- the presence of words in the reference lexica (TLFI and *lexique.org*) is beneficial to both precision and recall. We assume these lexica may represent common lexical competence of French speakers. For this reason, words that are present in these lexica, are also easier to understand;

- the frequencies of words computed through a general search engine are beneficial.

Words with higher frequencies are often associated with a better understanding, although the frequency range depends on the words. For instance, *coccyx* (*coccyx*) or *drain* (*drain*) show high frequencies (1,800,000 and 175,000,000, respectively) and they belong indeed to the *I can understand* category. Words like *colique* (*diarrhea*) or *clitoridien* (*clitoral*) show lower frequencies (807,000 and 9,821, respectively), although they belong to the same category. On contrary, other words with quite high frequencies, like *coagulase* (*coagulase*), *clivage* (*cleavage*) or *douve* (*fluke*) (655,000, 1,350,000 and 1,030,000, respectively) are not understood by the annotators.

According to these experiments, our results point out that, among the most efficient features, we can find syntactic categories, presence of words in the reference lexica, frequencies of words on Google and three- and four-character end substring. In comparison to the existing studies, such as those presented during the SemEval challenge (Specia et al., 2012), we propose to exploit a more complete set of features, several of which rely on the NLP methods (*e.g.*, syntactic tagging, morphological analysis). Especially the syntactic tagging appears to be salient for the task. In comparison to work done on general language data (Gala et al., 2013), our experiment shows better results (between 0.825 and 0.948 *accuracy* against 0.62 *accuracy* in the cited work), which indicates that specialized domains have indeed very specific words. Additional tests should be performed to obtain a more detailed impact of the features.

6.4 Usefulness of the method

We applied the proposed method to words from discharge summaries. The documents are preprocessed according to the same protocol and the words are assigned the same features as previously (section 5). The model learned on the *unanimity* set is applied. The results are shown in Figure 3. Among the words categorized as non-understandable (in red and underlined), we find:

- abbreviations (*NIHSS, OAP, NaCl, VNI*);
- technical medical terms (*hypoesthésie* (*hypoesthesia*), *parésie* (*paresia*), *thrombolyse* (*thrombolysis*), *iatrogène* (*iatrogenic*), *oxygénothérapie* (*oxygen therapy*), *désaturation* (*desaturation*));

Histoire de la maladie
Le patient a été hospitalisé le 18 / 7 / 11 à PELLEGRIN pour un AVC ischémique dans le territoire profond de l' artère cérébrale postérieure droite , thrombolysé à H + 3 .

Le patient présente , comme déficit , une hypoesthésie gauche et une parésie gauche (force motrice à 1 / 5 au membre supérieur gauche et 2 / 5 au membre inférieur gauche) , un NIHSS à 8 , une désorientation tempora-spatiale et une vigilance fluctuante .
Dans les suites , est survenu un OAP post thrombolyse , probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse) .

Le patient est donc transféré en réanimation : l' OAP est résolutif sous VNI et oxygénothérapie .

La majoration de l' insuffisance rénale nécessite 2 cures de dialyse . Mr K . est ensuite transféré en post-réanimation devant l' évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge .

Le 11 / 8 / 2011 , le patient présente une douleur thoracique associée à une désaturation à 83 % , il est donc transféré en Unité de soins intensifs cardiologiques . Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire . Une anticoagulation curative par CALCIPARINE est mise en place .

Figure 3: Detection of non-understandable words within discharge summaries.

- medication names (*CALCIPARINE*);

In the example from Figure 3, three types of errors can be distinguished when common words are categorized as non-understandable:

- inflected forms of words (*suites* (*consequences*), *cardiologiques* (*cardiological*));
- constructed forms of words (*thrombolysé* (*with thrombolysis*));
- hyphenated words (*post-réanimation* (*post emergency medical service*)).

Notice that in other processed documents, other errors occur. For instance, misspelled words and words that miss accented characters (*probleme* instead of *problème* (*problem*), *realise* instead of *réalisé* (*done*), *particularite* instead *particularité* (*particularity*)) are problematic. Another type of errors may occur when technical words (*e.g. prolapsus* (*prolapsus*), *paroxysme* (*paroxysm*), *tricuspide* (*tricuspid*)) are considered as understandable.

Besides, only isolated words are currently processed, which is the limitation of the current method. Still, consideration of complex medical terms, that convey more complex medical notions, should also be done. Such terms may indeed change the understanding of words, as in these examples: *AVC ischémique* (*ischemic CVA (cerebrovascular accident)*), *embolie pulmonaire basale droite* (*right basal pulmonary embolism*), *désaturation à 83 %*

(*desaturation at 83%*), *anticoagulation curative* (*curative anticoagulation*). In the same way, numerical values may also arise misunderstanding of medical information. Processing of these additional aspects (inflected and constructed forms of words, hyphenated or misspelled words, complex terms composed with several words and numerical values) is part of the future work.

6.5 Limitations of the current study

We proposed several experiments for analyzing the understandability of medical words. We tried to analyze these data from different points of view to get a more complete picture. Still, there are some limitations. These are mainly related to the linguistic data and to their preparation.

The whole set of the analyzed words is large: almost 30,000 entries. We assume it is possible that annotations provided may show some intra-annotator inconsistencies due for instance to the tiredness and instability of the annotators (for instance, when a given unknown morphological components is seen again and again, the meaning of this component may be deduced by the annotator). Nevertheless, in our daily life, we are also confronted to the medical language (our personal health or health of family or friend, TV and radio broadcast, various readings of newspapers and novels) and then, it is possible that the new medical notions may be learned during the annotation period of the words, which lasted up to four weeks. Nevertheless, the advantage of the data we have built is that the whole set is completely annotated by each annotator.

When computing the features of the words, we have favored those, which are computed at the word level. In the future work, it may be interesting to take into account features computed at the level of morphological components or of complex terms. The main question will be to decide how such features can be combined all together.

The annotators involved in the study have a training in linguistics, although their relation with the medical field is poor: they have no specific health problems and no expertise in medical terminology. We expect they may represent the average level of patients with moderate health literacy. Nevertheless, the observed results may remain specific to the category of young people with linguistic training. Additional experiments are required to study this aspect better.

7 Conclusion and Future research

We proposed a study of words from the medical field, which are manually annotated as understandable, non-understandable and possibly understandable to laymen. The proposed approach is based on machine learning and a set with 24 features. Among the features, which appear to be salient for the diagnosis of understandable words, we find for instance the presence of words in the reference lexica, their syntactic categories, their final substring, and their frequencies on the web. Several features and their combinations can be distinguished, which shows that the understandability of words is a complex notion, which involves several linguistic and extra-linguistic criteria.

The avenue for future research includes for instance the exploitation of corpora, while currently we use features computed out of context. We assume indeed that corpora may provide additional relevant information (semantic or statistical) for the task aimed in this study. Additional aspects related to the processing of documents (inflected and constructed forms of words, hyphenated or misspelled words, complex terms composed with several words and numerical values) is another perspective. Besides, the classical readability measures exploited have been developed for the processing of English language. Working with French-language data, we should use measures, which are adapted to this language (Kandel and Moles, 1958; Henry, 1975). In addition, we can also explore various perspectives, which appear from the current limitations, such as computing and using features computed at different levels (morphological components, words and complex terms), applying other classical readability measures adapted to the French language, and adding new reference annotations provided by laymen from other social-professional categories.

Acknowledgments

This work is performed under the grant ANR/DGA Tecsan (ANR-11-TECS-012) and the support of MESHS (COMETE project). The authors are thankful to the CHU de Bordeaux for making available the clinical documents.

References

AMA. 1999. Health literacy: report of the council on scientific affairs. Ad hoc committee on health lit-

eracy 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.

A Borst, A Gaudinat, C Boyer, and N Grabar. 2008. Lexically based distinction of readability levels of health documents. In *MIE 2008*. Poster.

MT Cabré and R Estopà. 2002. On the units of specialised meaning uses in professional com- munication. In *International Network for Terminology*, pages 217–237.

TM Cabré. 2000. Terminologie et linguistique: la thorie des portes. *Terminologies nouvelles*, 21:10–15.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

B Daille. 1995. Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *Traitement Automatique des Langues (T.A.L.)*, 36(1-2):101–118.

P Drouin and P Langlais. 2006. valuation du potentiel terminologique de candidats termes. In *JADT*, pages 379–388.

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

JL Fleiss and J Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

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

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

KT Frantzi, S Ananiadou, and J Tsujii. 1997. Automatic term recognition using contextual clues. In *MULSAIC IJCAI*, pages 73–79.

N Gala, T François, and C Fairon. 2013. Towards a french lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLEX-2013*.

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

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

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

G Henry. 1975. *Comment mesurer la lisibilit*. Labor, Bruxelles.

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

C Iacobini, 2003. *Composizione con elementi neoclassici*, pages 69–96.

Gonia Jarema, Cline Busson, Rossitza Nikolova, Kyrana Tsapkini, and Gary Libben. 1999. Processing compounds: A cross-linguistic study. *Brain and Language*, 68(1-2):362–369.

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.

R Jucks and R Bromme. 2007. Choice of words in doctor-patient communication: an analysis of health-related internet sites. *Health Commun*, 21(3):267–77.

L Kandel and A Moles. 1958. Application de lindice de flesch la langue franaise. *Cahiers tudes de Radio-Tlvision*, 19:253–274.

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

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.

I Korkontzelos, IP Klapaftis, and S Manandhar. 2008. Reviewing and evaluating automatic term recognition techniques. In *GoTAL*, pages 248–259.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

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.

Gary Libben, Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84(1):50–64.

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.

A Lüdeling, T Schmidt, and S Kiokpasoglou. 2002. Neoclassical word formation in german. *Yearbook of Morphology*, pages 253–283.

D Maynard and S Ananiadou. 2000. Identifying terms by their family and friends. In *Proceedings of COLING 2000*, pages 530–536, Saarbrucken, Germany.

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

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.

JR Quinlan. 1993. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

R Rittman. 2008. *Automatic discrimination of genres*. VDM, Saarbrucken, Germany.

R Rudd, B Moeykens, and T Colton, 1999. *Annual Review of Adult Learning and Literacy*, page ch 5.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 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.

I.H. Witten and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Eugen Wüster. 1981. L'tude scientifique gnrale de la terminologie, zone frontalire entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In G. Rondeau et H. Felber, editor, *Textes choisis de terminologie*, volume I. Fondements thoriques de la terminologie, pages 55–114. GISTERM, Universit de Laval, Qubec. sous la direction de V.I. Siforov.

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.

W Zheng, E Milios, and C Watters. 2002. Filtering for medical news items using a machine learning approach. In *AMIA*, pages 949–53.