Understanding of unknown 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, Université Paris-Saclay,  
Orsay, France  
Université Paris 13, Sorbonne Paris Cité,  
Villetaneuse, France  
hamon@limsi.fr

Abstract

We assume that unknown words with internal structure (affixed words or compounds) can provide speakers with linguistic cues as for their meaning, and thus help their decoding and understanding. To verify this hypothesis, we propose to work with a set of French medical words. These words are annotated by five annotators. Then, two kinds of analysis are performed: analysis of the evolution of understandable and non-understandable words (globally and according to some suffixes) and analysis of clusters created with unsupervised algorithms on basis of linguistic and extra-linguistic features of the studied words. Our results suggest that, according to linguistic sensitivity of annotators, technical words can be decoded and become understandable. As for the clusters, some of them distinguish between understandable and non-understandable words. Resources built in this work will be made freely available for the research purposes.

1 Introduction

Often, people face unknown words, be they neologisms (like in Some of the best effects in my garden have been the result of serendipity.) or technical words from specialized areas (like in Jacques Chirac’s historic corruption trial, due to start on Monday is on the verge of collapse, after doctors diagnosed him with “anosognosia”). In both cases, their semantics may be opaque and their understanding not obvious.

Several linguistic operations are available for enriching the lexicon, such as affixation, compounding and borrowings (Guilbert, 1971). We are particularly interested in words with internal structure, like anosognosia, because we assume that linguistic regularities (components, affixes, and rules that form their structure) can help speakers in deducing their structure and semantics. Our hypothesis is that if regularities can be observed at the level of linguistic features, they can also be deduced and managed by speakers. Indeed, linguistic understanding is related to factors like:

- knowledge and recognition of components of complex words: how to segment words, like anosognosia, in components;
- morphological patterns and relations between components: how to organize the components and to construct the word semantics (Jaubernini, 2003; Amiot and Dal, 2008).

To verify our hypothesis, we propose to work with a set of French medical words. These words are considered out of context for several reasons:

1. when new words appear, they have little and poor contexts, which cannot usually help their understanding;
2. similarly, in specialized areas, the contexts, except some definitional contexts, often bring little help for the understanding of terms;
3. working with words out of context permits to process a bigger set of words and to make observations with larger linguistic material;
4. from another point of view, analysis of words in context corresponds to their perception in extension relying on external clues, while analysis of words out of context corresponds to their perception in intension relying on clues and features internal to these words.

For these reasons, we assume that internal structure of unknown words can help their understanding. According to our hypothesis, affixed words
and compounds, which are given internal structure, can provide the required linguistic clues. Hence, the speakers may linguistically analyze unknown words thanks to the exploitation of their structure that they are able to detect.

Our interest for medical words is motivated by an increasing presence of medical notions in our daily life, while medicine still keeps a lot of mysteries unknown to lay persons because medical knowledge is typically encoded with technical and very specialized terms.

In what follows, we present some existing works (section 2), the data which we propose to process (section 3), and the experiments we propose to exploit (sections 4 to 6). We conclude with some orientations for future work (section 7).

2 Existing work

We concentrate on work related to text difficulty and understanding. Work on processing of words unknown in dictionaries by automatic applications, although well studied, is not presented.

NLP provides a great variety of work and approaches dedicated to understanding and readability of words and texts. The goal of readability is to define whether texts are accessible for readers or not. Readability measures are typically used for evaluation of document complexity. Classical readability measures exploit information on number of characters and syllables of words (Flesch, 1948; Gunning, 1973), while computational measures can involve vectorial models and different features, among which combination of classical measures with terminologies (Kokkinakis and Toporowska Gronostaj, 2006); n-grams of characters (Poprat et al., 2006); stylistic (Grabar et al., 2007) or discursive (Goeuriot et al., 2007) features; lexicon (Miller et al., 2007); morphological information (Chmielik and Grabar, 2011); and combination of various features (Wang, 2006; Zeng-Treiler et al., 2007; Leroy et al., 2008; François and Fairon, 2013; Gala et al., 2013).

In linguistics and psycholinguistics, the question on understanding of lexicon may focus on:

- Knowledge of components of complex words and their decomposition. The purpose is to study how complex words (affixed or compounds) are processed and recorded. Several factors may facilitate reading and production of complex words: when these compounds contain hyphens (Bertram et al., 2011) or spaces (Frisson et al., 2008); when they are presented with other morphologically related words (Lüttmann et al., 2011); and when primes (Bozic et al., 2007; Beyersmann et al., 2012), pictures (Dohmes et al., 2004; Koester and Schiller, 2011) or favorable contexts (Cain et al., 2009) are used;

- Order of components and variety of morphological patterns. Position of components (head or modifier) proved to be important for processing of complex words (Libben et al., 2003; Holle et al., 2010; Feldman and Soltano, 1999). The notions of semantic transparency and of morphological headedness have been isolated (Jarema et al., 1999; Libben et al., 2003);

- Word length and types of affixes (Meinzer et al., 2009);

- Frequency of bases and components (Feldman et al., 2004).

Our hypothesis on emerging of linguistic rules involved in word formation has also been addressed in psycholinguistics, and it has to face two other hypothesis on acquisition in context and on providing explicit information on semantics of components (Baumann et al., 2003; Kuo and Anderson, 2006; McCutchen et al., 2014). Currently, the importance of morphological structure for word processing seems to be accepted by psycholinguists (Bowers and Kirby, 2010), which supports our hypothesis. Yet, in our work, for verifying this hypothesis, we exploit NLP methods and NLP-generated features. Hence, we can work with large linguistic data and exploit quantitative and unsupervised methods.

3 Exploited data

The data processed are obtained from medical terminology Snomed International (Côté, 1996) in French, which purpose is to describe the medical area. This terminology contains 151,104 terms structured in eleven semantic axes (e.g. disorders and abnormalities, medical procedures, chemicals, leaving organisms, anatomy). We keep terms from five axes (disorders, abnormalities, medical procedures, functions and anatomy), which we consider to be central and frequent. Hence, we do not wish to concentrate on very specialized terms and
words, like chemicals or leaving organisms. Nevertheless, such words can be part of terms studied here. The selected terms (104,649) are segmented in words to obtain 29,641 unique words, which are our working material. This set contains compounds \textit{(abdominoplastie (abdominoplasty), dermabrasion (dermabrasion))}, constructed \textit{(cardiaque (cardiac), lipoïde (lipoid))} and simple \textit{(fragment)} words, as well as abbreviations \textit{(AD-Pase, ECQG, Fya)} and borrowings \textit{(stripping, Conidiobolus, stent, blind)}.

These terms are annotated by five French native speakers, aged from 25 to 60, without medical training and with different social and professional status. Each annotator received a set with randomly ordered 29,641 words. According to the guidelines, the annotators should not use additional information (dictionaries, encyclopedia, etc.), should not change annotations done previously, should manage their time and efforts, and assign each word in one of the three categories: (1) \textit{I can understand}, containing known words; (2) \textit{I am not sure}, containing hesitations; (3) \textit{I cannot understand}, containing unknown words.

We assume that our annotators represent moderate readability level \cite{Schonlau2011}, i.e. the annotators have a general language proficiency but no specific knowledge in medical domain, and that we will be able to generalize our observations on the same population. Besides, we assume that these annotations will allow to observe the progression in the understanding of technical words.

Manual annotation required from 3 weeks up to 3 months. The inter-annotator agreement \cite{Cohen1960} is over 0.730. Manual annotation allows to distinguish several types of words which are difficult to understand: (1) abbreviations \textit{(e.g., OG, VG, PAPS, j. bat, cp)}; (2) proper names \textit{(e.g., Gougerot, Sjögren, Bentall, Glasgow, Babinski, Barthel, Cockcroft)}, which are often part of terms meaning disorders and procedures; (3) medications; (4) several medical terms meaning disorders, exams and procedures. These are mainly compounds \textit{(e.g. antihémophile (antihaemophilus), sclérodermie (scleroderma), hydrolase (hydrolasis), tympanectomie (tympanectomy), synesthésie (synesthesia))}; (5) borrowings; (6) words related to human anatomy \textit{(e.g. cloacal (cloacal), nasopharyngé (nasopharyngal), mitral (mitral), diaphragmatique (diaphragmatic), inguinal (inguinal), érythème (erythema), maxillo-facial (maxillo-facial), mésentérique (mesenteric), mésentère (mesentry))}.

4 Experiments

We propose two experiments:

1. Study of understanding progression of words globally and according to some components (section 5);
2. Unsupervised classification of words, analysis of clusters and their comparison with manual annotations (section 6).

5 Progression in word understanding

Progression of word understanding corresponds to the rate of understandable and non-understandable words at a given moment $t$ for a given annotator. This permits to observe whether the annotators can become familiar with some components or morphological rules, and improve their understanding of words while the annotation is going on. This analysis is done on the whole set of words and on words with some components.

Figure 1 indicates the evolution of the three categories of words. The line corresponding to I cannot understand is in the upper part of the graphs, while the line I can understand is in the lower part. The category I am not sure is always at the bottom. We can distinguish the following tendencies:

- Annotators A2, A1 and especially A5 show the tendency to decrease the proportion of unknown words. We assume that they are becoming more familiar with some components and bases, and that they can better manage medical lexicon;
- Annotators A1, and in a lesser way A2 and A4, show the tendency to decrease the number of hesitation (category 2). Indeed, the proportion of these words decreases, while the proportion of words felt as known (category 1) increases. Later, the number of known words seems not to increase, except for A5. Besides, this learning effect is especially observable with the top 2,000 words and it mainly affects the transition of hesitation words to known words;
- For annotators A3 and A4, after a small increase of proportion of unknown words, this proportion remains stable. We assume that the annotation process of a large lexicon did not allow to gain in understanding of components of the processed technical words.

Figures 2 and 3 show the evolution of understanding of words ending with -ite (-itis) (meaning inflammation) and -tomie (-tomy) (meaning removal), respectively. We can see that A5 has difficulty to understand these words: the percentage of unknown words is increasing, while on the whole set of words (figure 1(e)) this annotator shows the opposite tendency, with the percentage of unknown words decreasing. Annotators A2 and A4 also have understanding difficulties with these words. Figures of other annotators suggest that they make progress in decoding and understanding of words in -ite and -tomie. They first show an improvement in understanding of these words, and later there is another small progression. On the basis of these observations, we can see that, according to types of words, to their linguistic features and to the sensitivity of annotators, it is possible to make progressive improvement in understanding of technical lexicon which a priori is unknown by speakers. As already noticed, we assume that linguistic regularities play an important role in improving of the understanding of new lexicon. We propose to observe now if such regularities can also be detected by unsupervised clustering algorithms.

6 Unsupervised classification of words

Unsupervised classification is performed with several algorithms implemented in Weka: SOM (Kohonen, 1989), Canopy (McCallum et al., 2000), Cobweb (Fisher, 1987), EM (Dempster et al., 1977), SimpleKMeans (Witten and Frank, 2005). Excepting SimpleKMeans and EM, it is not necessary to indicate the expected number of clusters. Each word is described with 23 linguistic and extra-linguistic features, which can be grouped in 8 classes (an excerpt is provided in Table 1):

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

- **Presence of words in reference lexica.** We exploit two French reference lexica: TLFi\(^1\) and lexique.org\(^2\). TLFi is a dictionary of the French language covering XIX and XX centuries, and contains almost 100,000 entries.

\(^1\)http://www.atilf.fr/
\(^2\)http://www.lexique.org/
Figure 2: Evolution of percentage of words ending with -ite in each category.

Figure 3: Evolution of percentage of words ending with -tomie in each category.

<table>
<thead>
<tr>
<th>lemma</th>
<th>POS</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$f_g$</th>
<th>$f_t$</th>
<th>$n_{ba}$</th>
<th>$n_{bs}$</th>
<th>initial</th>
<th>final</th>
<th>$n_{bc}$</th>
<th>$n_{bv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>alarme</td>
<td>N</td>
<td>+</td>
<td>+</td>
<td>73400000</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>ala,alar,alarm</td>
<td>rme,arme,larme</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>hépatite</td>
<td>N</td>
<td>+</td>
<td>+</td>
<td>15300000</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>hép,hépa,hépat</td>
<td>ite,tite,tomie</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>angiocholite</td>
<td>N</td>
<td>-</td>
<td>+</td>
<td>74700</td>
<td>12</td>
<td>1</td>
<td>5</td>
<td>ang,angi,angio</td>
<td>ite,lite,olite</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>desmodontose</td>
<td>N</td>
<td>+</td>
<td>-</td>
<td>2050</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>des,desm,desmo</td>
<td>ose,tomie,tose</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1: Excerpt with features: $POS$-tag, presence in reference lexica (TLFI $l_1$ and lexique.org $l_2$), frequency in search engine $f_g$ and terminology $f_t$, number of semantic axes $n_{ba}$, number of syllables $n_{bs}$, initial and final substrings (initial, final), number of consonants $n_{bc}$, number of vowels $n_{bv}$. 
The document discusses the creation of a lexical database, lexique.org, designed for psycholinguistic experiments. It contains over 135,000 entries, including inflectional forms. The document outlines the criteria for selecting the most understandable words:

- **Frequency of words through a non-specialized search engine.** For each word, the authors query the Google search engine to determine its frequency on the web. They assume that words with higher frequency are easier to understand.

- **Frequency of words in medical terminology.** The authors also compute the frequency of words in the medical terminology Snomed International.

- **Number and types of semantic categories associated to words.** They exploit the information on semantic axes of Snomed International and assume that words which occur in several axes are more central.

- **Length of words in number of characters and syllables.** For each word, they compute the number of characters and syllables, believing that longer words may be more difficult to understand.

- **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 decomposition of lemmas into bases and affixes, and provides semantic explanation of the analyzed lexemes.

- **Initial and final substrings.** They compute the initial and final substrings of different length, from three to five characters. This allows them to isolate some components and possibly the morphological head of words.

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

The authors perform experiments with three feature sets:

- $E_c$: the whole set with 23 features,
- $E_r$: set with features reduced to linguistic properties of words, such as POS-tag, number of syllables, initial and final substrings, which permits to take into account observations from psycholinguistics (Jarema et al., 1999; Libben et al., 2003; Meinzer et al., 2009),
- $E_f$: set with linguistic features and frequency collected with the search engine, which permits to consider other psycholinguistic observations (Feldman et al., 2004).

With SimpleKMeans and EM, the authors perform two series of experiments, in which the number of clusters is set to 1,000 and 2,000 (for almost 30,000 individuals to cluster). They aim to find linguistic regularities of words in clusters, according to the features exploited. More specifically, they want to observe whether the content of clusters is related to the understanding of words.

### Table 2: Generated clusters

<table>
<thead>
<tr>
<th>Features</th>
<th>SOM</th>
<th>Canopy</th>
<th>Cobweb</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_c$: Full set (23)</td>
<td>5</td>
<td>62</td>
<td>33853</td>
</tr>
<tr>
<td>$E_r$: Reduced set (8)</td>
<td>4</td>
<td>28</td>
<td>12577</td>
</tr>
<tr>
<td>$E_f$: $E_r$ and frequency (9)</td>
<td>4</td>
<td>27</td>
<td>9861</td>
</tr>
</tbody>
</table>

In Table 2, we indicate the number of clusters obtained with various sets of features. SOM generates very few clusters, which are big and heterogeneous. For instance, with $E_f$, clusters contain up to 13,088, 4,840, 7,023 and 4,690 individuals; Cobweb generates a lot of clusters among which several singletons. For instance, with $E_f$, we obtain 9,374 clusters out of which 9,861 are singletons; EM and SimpleKMeans generate the required number of clusters, 1,000 and 2,000; Canopy generates between 30 and 60 clusters, according to the features used. We propose to work with clusters obtained with Canopy because it generates reasonable number of clusters, which number and contents are motivated by features.

With features from sets $E_r$ and $E_f$, cluster creation is mainly motivated by initial substrings (not always equal to 3 to 5 first or final characters) and in a lesser way by their POS-tags and frequencies. For instance, we can obtain clusters with words beginning by $p$ or $a$, or clusters grouping...
phosphats or enzymes ending with -ase. In this last case, clusters with chemicals become interesting for our purpose, although globally the clusters generated on basis of features from sets $E_r$ and $E_f$ show little interest. We propose to work with clusters obtained with the $E_c$ featureset.

With Canopy, the size of clusters varies between 1 and 2,823 individuals. Several clusters are dedicated to two main annotation categories. Hence, 30 clusters contain at least 80% of words from the category 1 (I can understand), while 6 clusters contain at least 80% of words from the category 3 (I cannot understand). Among the clusters with understandable words, we can find clusters with:

- numerals (mil (thousand), quinzième (fifteen)), verbs (allaite (breast-feed), étend (expand)), and adverbs (massivement (massively), probablement (probably)) grouped according to their POS-tags and sometimes to their final substrings;
- grammatical words (du (of), aucun (any), les (the)) grouped on basis of length and POS-tags;
- common adjectives (rudimentaire (rudimentary), prolongé (extended), perméable (permeable), hystérique (hysterical), inadéquat (inadequate), traumatique (traumatic), militaire (military)) grouped according to their POS-tags and frequency;
- participial adjectives (inapproprié (inappropriate), stratifié (stratified), relié (related), modifié (modified), localisé (localised), précisé (precise), quadruplé (quadrupled)) grouped according to their POS-tags, frequencies and final substrings;
- specialized but frequent adjectives (rotulien (patellar), spasmodique (spasmodic), putréfié (putrefactive), redondant (redundant), tremblant (trembling), vénal (venal), synchrone (synchronous), sensoriel (sensory)), also grouped according to their POS-tags and frequencies;
- specialized frequent nouns (dentiste (dentist), Brosse (brush), altitude (altitude), glucose (glucose), fourrure (fur), ankylose (ankylosis), aversion (aversion), carcinome (carcinoma)) grouped according to their POS-tags and frequencies.

Among the clusters with non-understandable words, we can find:

- chemicals (dihydroxyisovaléraste, héparosane-N-sulfate-glucuronate, désoxythymidine-monophosphate, diméthylallyltransférase) grouped according to their POS-tags, types of characters they contain and their frequency;
- borrowings (punctum, Saprolegnia, pig-mentsosum, framboesia, équuli, rubidium, dissimilis, frutescens, lontiasis, materia, méragonulc, diminutus, ghost, immitis, fol-llicis, musculi) grouped according to their POS-tags, final substrings and frequency;
- proper names grouped according to their POS-tags.

Within clusters with over 80% of words from the category 3 (I cannot understand), we do not observe understanding progression of annotators. Yet, we have several mixed clusters, that contain words from the two main categories (1 (I can understand) and 3 (I cannot understand)), as well as hesitations. These clusters contain for instance:

- chemicals and food (créatinine (creatine), antitussif (antitussive), céphalosporine (cephalosporine), aubergine (eggplant), carotte (carrot), antidépresseur (antidepressant), dioxide (dioxide)) grouped according to their final substrings, semantic axes and frequency;
- organism functions, disorders and medical procedures (paraparésie (paraparesis), névralgie (neuralgia), extrasystole (extrasystole), myéloblaste (myeloblast), syncope (syncpe), psychose (psychosis), spasticité (spasticity)) grouped according to their frequency, final substrings and POS-tags;
- more specialized adjectives related to anatomy and disorders (périprostatique (periprostatic), sous-entoriel (tensor), condylien (condylar), fibrosante (fibrotic), nécrosant (necrosis)) grouped according to their POS-tags and frequency.
Evolution of understanding is observable mainly within this last set of clusters. For instance, a typical example is the cluster containing medical procedures ending in -tomie, which words become less frequently assigned to the category 3 (I cannot understand) and more frequently to the categories 2 (I am not sure) and 1 (I can understand).

The content of clusters and our observations suggest that, given an appropriate set of features and unsupervised algorithms, it is possible to create clusters which reflect the readability and understandability of lexicon by lay persons. Besides, within some clusters, it is possible to observe the evolution of annotators in their understanding of technical words. For instance, this effect can typically be observed with words meaning disorders and procedures. Nevertheless, with other types of words (chemicals, borrowings, proper names) no evolution is observable.

Notice that the same reference data have been used with supervised categorization algorithms. In this case, automatic algorithms can reproduce the reference categorization with F-measure over 0.80 and up to 0.90, which is higher than the inter-annotator agreement rate. Besides, in the supervised categorization task, the behaviour of features is different from what we can observe in unsupervised clusters: several individual features can reproduce the reference categories while the best results are obtained with the whole set of features.

7 Conclusion and Future work

According to our hypothesis, linguistic regularities, when they occur systematically, can help in decoding and understanding of technical words with internal structure (like compounds or derived words). To test the hypothesis, we work with French medical words. Almost 30,000 words are annotated by five annotators and assigned in one of the three categories I can understand, I am not sure, I cannot understand. For each annotator, the words are ordered randomly.

We then perform an analysis of the whole set of words, and of words ending with -ite and -tomie. Our results suggest that several annotators show the learning effect as the annotation is going on, which supports our hypothesis and the findings of psycholinguistic work (Lüttmann et al., 2011). This effect is observed for the whole set of words and for the two analyzed suffixes. Yet, with chemicals, borrowings and proper names, we do not observe the learning effect.

These observations have been corroborated with clusters generated using linguistic and extralinguistic features. Several clusters are dedicated to words from either 1 (I can understand) or 3 (I cannot understand) categories. Besides, when clusters contain some semantically homogeneous words (disorders, procedures...) we can observe the expected learning effect. These results are very interesting and confirm our hypothesis, according to which linguistic regularities can help to decode and understand technical and unknown words. Appropriate features can also help to distinguish between understandable and non-understandable words with unsupervised methods. Correlations between social and demographic status and understanding require additional annotations. It will be studied in the future.

We have several directions for future work: (1) collect the same type of annotations, but providing semantics of some or of all components, although it will be difficult to verify whether this information is really exploited by annotators; (2) collect the same type of annotations, but permitting the annotators to use external sources of informations (dictionaries, online examples...). Since this approach requires more time and cognitive effort, smaller set of words will be used; (3) analyze the evolution of understanding of words taking into account a larger set of components; (4) validate the observations with tests for statistical significance; (5) exploit the results for training and education of non-experts in order to help them with the understanding of medical notions; (6) exploit the results for simplification of technical texts. For instance, features of words that show understanding difficulties can be used to define classes of words that should be systematically simplified.

The resources built in this work are freely available for the research purposes: http://natalia.grabar.free.fr/resources.php#rated.

Acknowledgments

We would like to thank the Annotators for their hard annotation work. This research has received aid from the IReSP financing partner within the 2016 general project call, Health service axis (grant GAGNAYRE-AAP16-HSR-6).
References


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*.


