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 (Iacobini, 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

Proceedings of the Biomedical NLP Workshop associated with RANLP 2017, pages 32–41,
Varna, Bulgaria, 8 September 2017.
https://doi.org/10.26615/978-954-452-044-1_005

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);

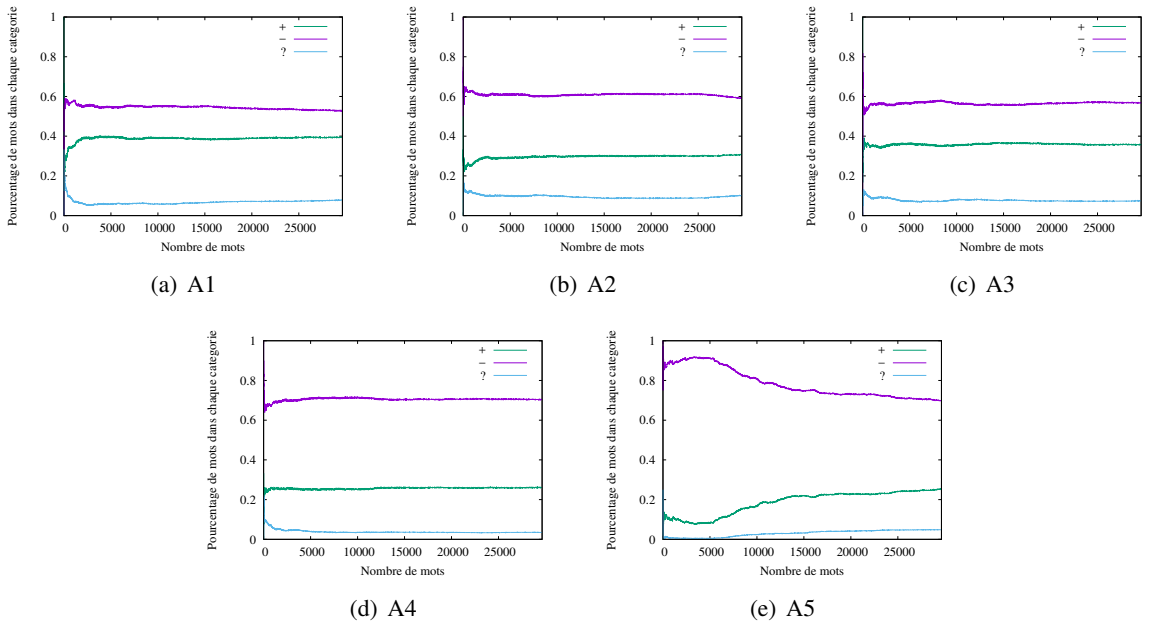
- Word length and types of affixes (Meinzer et al., 2009);

- 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



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

Figure 1: Global evolution of percentage of words per caterogy.

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 (*abdominoplastie (abdominoplasty), dermabrasion (dermabrasion)*), constructed (*cardiaque (cardiac), lipoïde (lipoid)*) and simple (*fragment*) words, as well as abbreviations (*ADPase, ECoG, Fya*) and borrowings (*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) *I can understand*, containing known words; (2) *I am not sure*, containing hesitations; (3) *I cannot understand*, containing unknown words. We assume that our annotators represent moderate readability level (Schonlau et al., 2011), 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 (Cohen, 1960) is over 0.730. Manual annotation allows to distinguish several types of words which are difficult to understand: (1) abbreviations (e.g. , *OG, VG, PAPS, j, bat, cp*); (2) proper names (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 (e.g. *antihémophile (anti haemophilus), sclérodermie (sclerodermia), hydrolase (hydrolasis), tympanectomie (tympanectomia), synesthésie (synesthesia)*); (5) borrowings; (6) words related to human anatomy (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.

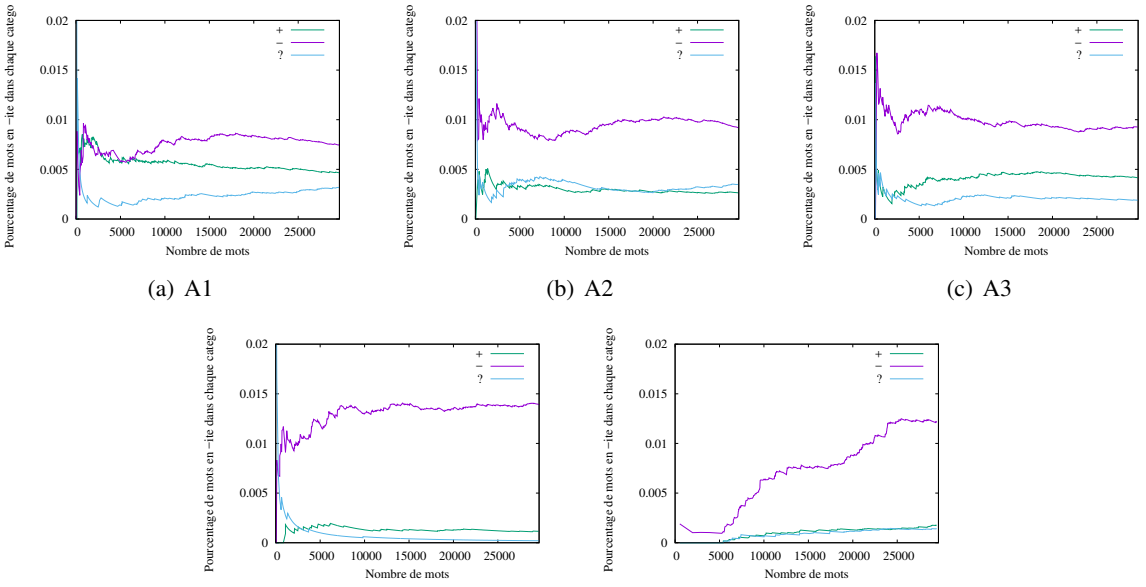
Figures 2 and 3 show the evolution of understanding of words ending with *-ite (-itis)* (meaning *inflammation*) and *-tomie (-tomy)* (meaning *removal*), respectively. We can see that A5 has difficulty to understand these words: the percentage of unknown words is increasing, while on the whole set of 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 it 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.

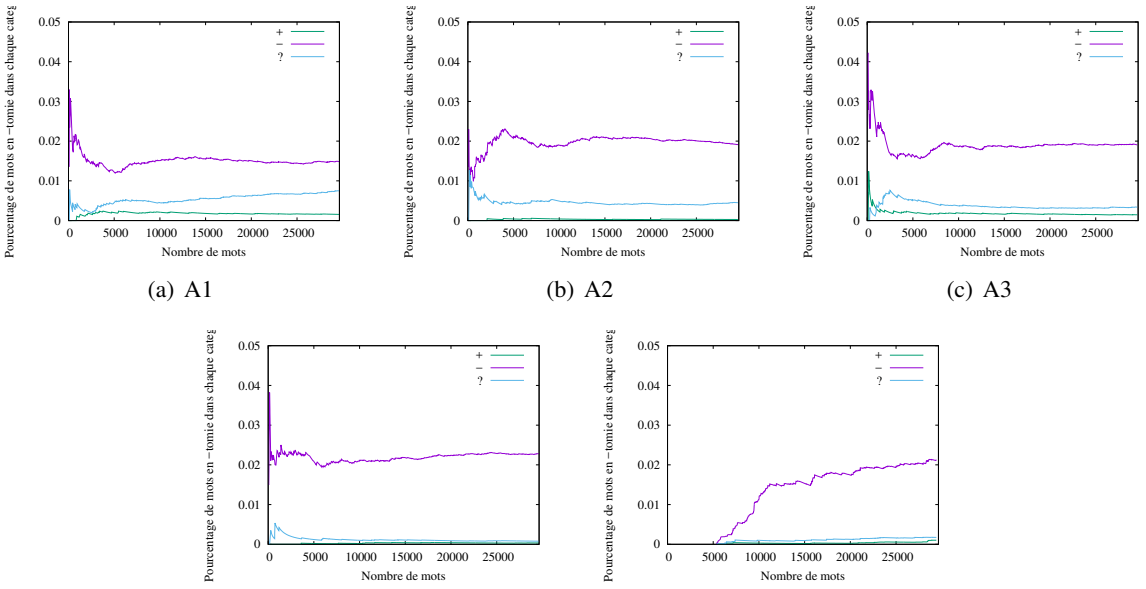
[1] http://www.atilf.fr/
[2] http://www.lexique.org/



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

Figure 2: Evolution of percentage of words ending with *-ite* in each category.



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

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

lemma	POS	l_1	l_2	f_g	f_t	nb_a	nb_s	initial	final	nb_c	nb_v
alarme	N	+	+	73400000	6	1	2	ala,alar,alarm	rme,arme,larme	3	3
hépatite	N	+	+	15300000	9	3	3	hép,hépa,hépat	ite,tite,atite	4	4
angiocholite	N	-	+	74700	12	1	5	ang,angi,angio	ite,lite,olite	6	6
desmodontose	N	+	-	2050	12	1	4	des,desm,desmo	ose,tose,ntose	7	5

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 nb_a, number of syllables nb_s, initial and final substrings (*initial, final*), number of consonants nb_c, number of vowels nb_v.

lexique.org has been created for psycholinguistic experiments. It contains over 135,000 entries, including inflectional forms. It contains almost 35,000 lemmas. We assume that words that are part of these lexica may be easier to understand;

- *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. We assume that words with higher frequency may be easier to understand;
- *Frequency of words in medical terminology.* For the same reason as above, we compute the frequency of words in the medical terminology Snomed International;
- *Number and types of semantic categories associated to words.* We also 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, we compute the number of its characters and syllables, because we think 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. We exploit morphological decomposition, which permits to compute the number of affixes and bases. Here again we focus on complexity of the internal structure of words;
- *Initial and final substrings.* We compute the initial and final substrings of different length, from three to five characters. This allows to isolate some components and possibly the morphological head of words;
- *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).

We perform experiments with three featuresets:

- 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`, we 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). We expect to find linguistic regularities of words in clusters, according to the features exploited. More specifically, we want to observe whether the content of clusters is related to the understanding of words.

Features	*SOM*	*Canopy*	*Cobweb*
E_c: Full set (23)	5	62	33853
E_r: Reduced set (8)	4	28	12577
E_f: E_r and frequency (9)	4	27	9861

Table 2: Generated clusters

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 reasonnable 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 (*dihydroxyisovalérate, héparosane-N-sulfate-glucuronate, désoxythymidine-monophosphate, diméthylallyltransférase*) grouped according to their POS-tags, types of characters they contain and their frequency;
- borrowings (*punctum, Saprolegnia, pigmentosum, framboesia, equuli, rubidium, dissimilis, frutescens, léontiasis, materia, mégarectum, diminutus, ghost, immitis, folliclis, 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 (creatinine), antitussif (antitussive), céphalosporine (cephalosporine), aubergine (eggplant), carotte (carrot), antidépresseur (antidepressant), dioxyde (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 (syncope), 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-tentoriel (tensor), condylienne (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 extra-linguistic 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

D Amiot and G Dal. 2008. La composition néoclassique en français et ordre des constituants. *La composition dans les langues* pages 89–113.

JF Baumann, EC Edwards, EM Boland, S Olejnik, and EJ Kame'enui. 2003. Vocabulary tricks: Effects of instruction in morphology and context on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal* 40(2):447–494.

Raymond Bertram, Victor Kuperman, Harald R Baayen, and Jukka Hyönä. 2011. The hyphen as a segmentation cue in triconstituent compound processing: It's getting better all the time. *Scandinavian Journal of Psychology* 52(6):530–544.

Elisabeth Beyersmann, Max Coltheart, and Anne Castles. 2012. Parallel processing of whole words and morphemes in visual word recognition. *The Quarterly Journal of Experimental Psychology* 65(9):1798–1819.

PN Bowers and JR Kirby. 2010. Effects of morphological instruction on vocabulary acquisition. *Reading and Writing* 23(5):515–537.

Mirjana Bozic, William D. Marslen-Wilson, Emmanuel A. Stamatakis, Matthew H. Davis, and Lorraine K. Tyler. 2007. Differentiating morphology, form, and meaning: Neural correlates of morphological complexity. *Journal of Cognitive Neuroscience* 19(9):1464–1475.

Kate Cain, Andrea S. Towse, and Rachael S. Knight. 2009. The development of idiom comprehension: An investigation of semantic and contextual processing skills. *Journal of Experimental Child Psychology* 102(3):280–298.

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.

J 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.

AP Dempster, NM Laird, and DB Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* 39(1):1–38.

Petra Dohmes, Pienie Zwitserlood, and Jens Bölte. 2004. The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language* 90(1-3):203–212.

Laurie Beth Feldman and Emily G. Soltano. 1999. Morphological priming: The role of prime duration, semantic transparency, and affix position. *Brain and Language* 68(1-2):33–39.

Laurie Beth Feldman, Emily G Soltano, Matthew J Pastizzo, and Sarah E Francis. 2004. What do graded effects of semantic transparency reveal about morphological processing? *Brain and Language* 90(1-3):17–30.

Douglas Fisher. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2(2):139–172.

R Flesch. 1948. A new readability yardstick. *Journ Appl Psychol* 23:221–233.

T François and C Fairon. 2013. Les apports du TAL à la lisibilité du français langue étrangère. *TAL* 54(1):171–202.

S Frisson, E Niswander-Klement, and A Pollatsek. 2008. The role of semantic transparency in the processing of english compound words. *Br J Psychol* 99(1):87–107.

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.

L Guilbert. 1971. De la formation des unités lexicales. In Paris Larousse, editor, *Grand Larousse de la langue française*, pages IX–LXXXI.

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

Henning Holle, Thomas C Gunter, and Dirk Koester. 2010. The time course of lexical access in morphologically complex words. *Neuroreport* 21(5):319–323.

C Iacobini. 2003. Composizione con elementi neoclassici. In Maria Grossmann and Franz Rainer, editors, *La formazione delle parole in italiano*, Walter de Gruyter, pages 69–96.

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

Dirk Koester and Niels O. Schiller. 2011. The functional neuroanatomy of morphology in language production. *NeuroImage* 55(2):732–741.

T Kohonen. 1989. *Self-Organization and Associative Memory*. Springer.

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.

LJ Kuo and RC Anderson. 2006. Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist* 41(3):161–180.

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.

Heidi Lüttmann, Pienie Zwitserlood, and Jens Bölte. 2011. Sharing morphemes without sharing meaning: Production and comprehension of german verbs in the context of morphological relatives. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 65(3):173–191.

A McCallum, K Nigam, and LH Ungar. 2000. Efficient clustering of high dimensional data sets with application to reference matching. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. pages 169–178.

Deborah McCutchen, Sara Stull, Becky Logan Herrera, Sasha Lotas, and Sarah Evans. 2014. Putting words to work: Effects of morphological instruction on children's writing. *J Learn Disabil* 47(1):1–23.

Marcus Meinzer, Aditi Lahiri, Tobias Flaisch, Ronny Hannemann, and Carsten Eulitz. 2009. Opaque for the reader but transparent for the brain: Neural signatures of morphological complexity. *Neuropsychologia* 47(8-9):1964–1971.

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.

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

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.

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*. Maastricht, pages 503–508.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. pages 44–49.

M Schonlau, L Martin, A Haas, KP Derose, and R Rudd. 2011. Patients' literacy skills: more than just reading ability. *J Health Commun* 16(10):1046–54.

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.

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*. Brisbane, Australia, pages 1117–1121.