The comprehension of medical words: Cross-lingual experiments in French and Xhosa

Natalia Grabar[1], Izak van Zyl[2], Retha de la Harpe[2] and Thierry Hamon[3]
[1]*CNRS UMR 8163 STL, Université Lille 1&3, 59653 Villeneuve d'Ascq, France*
[2]*Cape Peninsula University of Technology, Cape Town, South Africa*
[3]*LIMSI-CNRS, BP133, Orsay, France; Université Paris 13, Sorbonne Paris Cité, France*
natalia.grabar@univ-lille3.fr, izak.vzyl@gmail.com, retha.delaharpe@gmail.com, thierry.hamon@limsi.fr

Keywords: Health Literacy, Readability, Consumer Health Informatics, Natural Language Processing, Xhosa, French.

Abstract: This paper presents cross-lingual experiments in automatic detection of medical words that may be difficult to understand by patients. The study relies on Natural Language Processing (NLP) methods, conducted in three steps, across two languages, French and Xhosa: (1) the French data are processed by NLP methods and tools to reproduce the manual categorization of words as understandable or not; (2) the Xhosa data are clustered with a non-supervised algorithm; (3) an analysis of the Xhosa results and their comparison with the results observed on the French data is performed. Some similarities between the two languages are observed.

1 INTRODUCTION

Health information is integral to individual lifeways in light of pressing health concerns and the pursuit of maintaining healthy lifestyles. Moreover, health information is already widespread in society and is disseminated through a variety of media, scientific and medical research, news, the internet, radio and TV programs. However, the availability of this information does not guarantee its correct understanding and use. Standard medical and health language indeed conveys very specialized and technical notions elaborated and usually disseminated by healthcare professionals. These notions often remain opaque and non understandable for non-expert users, and especially for patients (Berland et al., 2001). This is despite the fact that they may have an important influence on the success of patients' medical care and their quality of life. For such reasons, the difficulty related to the access and use of health information must be addressed and surmounted. This extends to a process of enhanced communication between medical professionals and patients; a better understanding of the medical care delivered to patients; better management of chronic diseases (AMA, 1999).

Another aspect is related to multilingual and multicultural contexts in societies and in hospitals. Typically, health care providers operate within distinct language and cultural modalities, which may often diverge from the patient context. This disparity usually impedes communication leading to patient dissatisfaction and potential misdiagnosis and inappropriate medication prescriptions. Existing studies have addressed, inter alia, Spanish-speaking communities in North America (Woloshin et al., 1995; Flores et al., 1998), and Xhosa-speaking communities in South Africa (Levin, 2006b; Levin, 2006a; Schlemmer and Mash, 2006). Miscommunication and language disparities reveal additional complexities as they are embedded within cultural, lingual, and specialized and technical aspects of communication.

The possibility of simplified health information use has been poorly researched up to now in the Computer Sciences field. The studies of the kind are done in connection within the Consumer Health Informatics. They join at least two aspects: (1) Semantic Interoperability (through study and creation of relations between expert and non-expert languages in order to improve the communication), and (2) Computer Sciences and Disability, because Computer Sciences provide methods for helping patients with understanding the health information (Zeng and Parmanto, 2003). Other studies address the automatic distinction between specialized and non-specialized documents and words (Miller et al., 2007; François and Fairon, 2013), and focus on the building of expert and non-expert aligned medical vocabularies (Zeng et al., 2006; Elhadad and Sutaria, 2007; Deléger and

Zweigenbaum, 2008) in one given language (*i.e.*, English and French in the cited studies), or on the building of bilingual patient-oriented French-English terminology (Messai et al., 2006). Also, a specific task has been proposed withinin the SemEval challenge aiming the substitution of words by their simpler equivalents in general language texts in English (Specia et al., 2012). In the present study, we propose to focus on the distinction between comprehensible and non-comprehensible words from the medical field in two languages from two distant families, Xhosa and French. We rely on the Computer Sciences and on Natural Language Processing (NLP) methods and tools. The main hypothesis is that, when working with different languages and applying different methods and resources, it remains possible to observe some similarities between the languages.

In the following of this study, we first present the background and objectives of our study (section 2), we then describe the material used (section 3), and the methods we propose to achieve the objectives (section 4). We then discuss the obtained results (section 5), and present conclusions (section 6).

2 BACKGROUND

2.1 Specificities of the two studied languages

We indicate here some specificities and differences between Xhosa and French. *Xhosa* is a Bantu language, spoken in South Africa. It is a tonal language, which means that the same sequence of consonants and vowels can have different meanings when said with a rising or falling or high or low intonation. Tone is contrastive in Xhosa, with several examples of words that differ only in tone (Van der Stouwe, 2009): *goba* can mean *be happy* or *put lotion on*, and *imithi* can mean *its pregnant* or *trees* according to tones. Furthermore, Xhosa is an agglutinative language, which means that prefixes and suffixes are attached directly to root words. For instance, in *abantwana bayadlala*, the noun *ntwana* (*child*) is modified by the plural affix *aba-*, while the verb *dlala* (*play*) is modified by the affix *baya-* meaning third plural person in present tense. The Xhosa words belong mainly to two syntactic classes (nouns and verbs), while the affixes (prefixes and suffixes) convey additional meanings, like in the presented example. From the point of view of NLP, the Xhosa language has recently attracted attention of researchers (Allwood et al., 2003; Roux et al., 2004; Moropa, 2007; Bosch et al., 2008; Pretorius and Bosch, 2009), but up to now there is a paucity of freely available resources and tools. *French* is a Latin language, spoken in Europe and in countries across the world. French grammar organizes words into several syntactic classes (nouns, verbs, adjectives, adverbs, prepositions...). Its words can be modified with inflectional ({*abdomen*, *abdomens*}), derivational ({*abdomen*, *abdominal*}) and compositional ({*abdomen*, *abdominoplastie*}) processes. French is not an agglutinative language: words and syntactic classes remain free in the sentences, although they have defined places there. The semantic ambiguity of words is mainly due to the inherent semantics of their morphological bases. For instance, *montre* means *a watch* and some inflectional forms of the verb *show*; while *poste* means *post office* and *employment position*. The French language has been the object of several NLP research studies and is provided with several resources (dictionaries, lexica, terminologies, etc.) and tools.

2.2 Rationale of the study

Given the differences that exist between the two studied languages, it is difficult to apply the same methods for processing data from each. For this reason, the French data are processed with supervised methods and a sophisticated set of linguistic features involving several resources and NLP tools (section 4.1), while the Xhosa data are processed with unsupervised methods and poorer linguistic data (section 4.2). In Table 1, we summarize the rationale of the study: in Xhosa, the unsupervised methods are applied to the surface forms of words which linguistic description relies on the contexts of these words; while in French, the supervised methods are applied to the lemmas of words which are provided with a rich linguistic description.

Table 1: Rationale of the methods for Xhosa and French.

	Xhosa	French
Methods	unsupervised	supervised
Units	forms	lemmas
Features	contexts	rich set of features

2.3 Objectives

We have several objectives: (1) analyze how the distinction between technical and understandable words can be done automatically; (2) work with multilingual data from the medical field; (3) apply different methods to the different languages but obtain comparable results; (4) study how and to which extent we can enrich the linguistic description of the less resourced language.

3 MATERIAL

We use two datasets related to the medical field language, one in French and one in Xhosa. We also use two lexica (English and bilingual Xhosa-English).

3.1 Building and preprocessing the French linguistic data

In French, the data are obtained from the medical terminology Snomed International (Côté, 1996), in its version currently distributed by ASIP Santé[1]. We use this terminology because its aim is to describe the entire medical field: this provides us with the possibility to study the main medical notions extensively. Besides, the terms recorded in Snomed Int are usually extracted from real clinical documents and often correspond to real expressions used in the language, and with which the patients may be faced.

Snomed contains 151,104 terms structured into eleven semantic axes. Among these, we study five taxonomies related to the main medical notions (disorders, abnormalities, procedures, functions, and anatomy). The corresponding terms (104,649) are tokenized and segmented into words. We obtain 29,641 unique words: they correspond to our main material and we use them in order to study how the complexity of specialized words is felt by the speakers and how it can be described and processed with automatic approaches. The preprocessing of words consists of automatic assignment to a syntactic category (Schmid, 1994; Namer, 2000), such as noun (*agnosie, adénome*), adjective (*bacteroide, D-dépendant, abactérien*), verb (*alimenter, baver, changer*), adverb (*essentiellement, crescendo, facilement*), preposition or determinant; and their lemmatization that provides canonical form of words (singular for nouns, masculine singular for adjectives, infinitive for verbs). In the following, all the experiments in French are performed on lemmas. These linguistic data are also annotated according to the aims of our study: three speakers without any medical training, considered as laymen, are involved. They are asked to analyze the 29,641 words and to assign them to one of the three categories:

1. *I can understand the word*;
2. *I am not sure about its meaning*;
3. *I cannot understand the word*.

The assumption is that the words that cannot be understood by the annotators should be considered as semantically difficult. These manual annotations correspond to the reference data.

[1] http://esante.gouv.fr/asip-sante

3.2 Building and preprocessing the Xhosa linguistic data

In Xhosa, the linguistic data are mainly obtained from brochures created for patients (prevention, pathology and treatments for HIV, rape and other medical conditions). They are freely available and are collected from several websites[2]. These documents are converted from PDF to text format to make NLP processing possible. The final corpus contains 34 documents and over 206,000 tokens of words. This corpus corresponds to the main material exploited in Xhosa. The corpus is further preprocessed:

- tokenizing the punctuation, *i.e.* separation of dots, commas, which is the first step of normalization;
- recognition and filtering out sentences in English, as in the exploited documents some sentences in English may occur.

3.3 Lexica

The lexicon of English words is built using the Robert & Collins dictionary. It contains 398,229 entries. It is used to filter out sentences in English. A bilingual Xhosa-English lexicon is built to help interpret the results generated from the Xhosa corpus. It is built using the data available on different websites[3], and from the inventory of Xhosa plant names (Dold and Cocks, 1999). It is also completed with word pairs[4] while analyzing the clusters. The resulting lexicon contains 6,594 entries, including agglutinated forms.

4 METHODS

The proposed method is composed of three steps: (1) processing of the French data with supervised methods and choice of the descriptors suitable for the automatic distinction between understandable and non-understandable words; (2) exploitation of the Xhosa corpus with unsupervised methods and building of clusters of words; (3) interpretation of clusters.

4.1 Processing the French data

In French, the task is addressed as classification problem for the automatic distinction between understand-

[2] www.tac.org.za, www.doh.gov.za, www.capetown.gov.za

[3] www.travlang.com, sabelo.tripod.com/dictionary.htm, http://www.dicts.info

[4] http://mokennon.albion.edu/

able and non-understandable words. Supervised machine learning method is applied: the annotations produced by the annotators are used as training data for the creation of specific models, and as reference data for evaluation. This process relies on specific set of features (section 4.1.1), machine learning (section 4.1.2), and evaluation protocol (section 4.1.3).

4.1.1 Generation of the features

We exploit 24 features computed automatically. These features can be grouped into ten classes.

Syntactic categories. Syntactic categories and lemmas are computed by TreeTagger (Schmid, 1994) and 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. Among the main categories we find nouns, adjectives, proper names, verbs and abbreviations.

Presence of words in reference lexica. We exploit two reference lexica of the French language: TLFi (TLFi, 2001) and lexique.org[5]. TLFi is a dictionary of the French language covering XIX and XX centuries, with almost 100,000 entries. lexique.org is a lexicon created for psycholinguistic experiments, with over 135,000 entries, among which almost 35,000 lemmas. These two lexica are expected to represent the common lexical competence of speakers and we suppose that those words that are present in these lexica should be more familiar.

Frequency of words through a non-specialized search engine. We query a non-specialized search engine in order to know its frequency attested on the web. Those words that are more frequent are expected to be easier to understand.

Frequency of words in medical terminology. We compute the frequency of words in the medical terminology, Snomed International. Similarly, we suppose that words that are more frequent there can be less difficult to understand by layman speakers.

Number and types of semantic categories associated to words. We exploit information on the semantic categories of Snomed International: we expect that words that belong to several categories may convey more fundamental medical notions and be better known by speakers.

Length of words in number of their characters and syllables. We compute the number of characters and syllables, and expect that longer words are potentially more difficult to understand, because they can correspond to lexically complex lexemes.

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

Number of bases and affixes. Lemmas are analyzed by the morphological analyzer Dérif (Namer, 2009). It performs their decomposition into bases and affixes known in its database. Here again, we expect that morphologically more complex lemmas may correspond to semantically more complex lexemes.

Initial and final substrings of the words. We compute the initial and final substrings (three to five characters). We expect that these substrings may be evocative of the bases or affixes positioned at the beginning and, especially, at the end of words. The main motivation is that final substrings correspond to the semantic base of compounds, often Latin of Greek components.

Number and percentage of consonants, vowels and other characters. We also compute the number and the percentage of consonants, vowels and other characters (for instance, hyphen, apostrophe, comas such as they occur in names of chemical products).

Classical readability scores. We apply two classical readability measures: Flesch (Flesch, 1948) and its variant Flesch-Kincaid (Kincaid et al., 1975). Typically used for evaluating the difficulty level of texts, they exploit surface characteristics of words (number of characters and/or syllables) and normalize these values with specific coefficients. Longer words are considered to be more difficult to understand.

4.1.2 Machine learning system

Table 2: Number (and percentage) of words assigned to reference categories within the majority set.

Categories	*Number*	*%*
1. I can understand	7,655	27
2. I am not sure	597	2
3. I cannot understand	20,511	71
Total annotations	28,763	100

Machine learning is used in order to classify the data and to distinguish between comprehensible words among laymen, and also to study the importance of various features for the task. The machine learning exploits an annotated dataset, that is described with suitable features such as those presented above. On the basis of such features, the algorithms can detect the regularities within the training dataset to generate a model and apply the generated model to process new unseen data. We apply several algorithms available in WEKA (Witten and Frank, 2005).

The annotations, provided by the three annotators, constitute our reference data. We use here the dataset *majority* (Table 2) that contains the annotations for which we can compute the majority agreement of the annotators *i.e.*, at least two of the annotators agree.

This dataset contains 28,763 words (out of 29,641), among which 71% are assigned to *I cannot understand*, 27% to *I can understand* and only 2% to *I am not sure*: the non-comprehensible words are the most frequent. According to the Fleiss' Kappa (Fleiss and Cohen, 1973), suitable for processing the data provided by more that two annotators, the inter-annotator agreement shows substantial agreement (Landis and Koch, 1977), with the score 0.73. This corresponds to a very good agreement level, especially when working with linguistic data, for which the agreement is usually difficult to obtain, as it greatly depends on the individual linguistic feeling of the speakers.

4.1.3 Evaluation

The success of the machine learning algorithms is evaluated with three classical measures: recall $\mathcal{R}$ (how exhaustive are the results?), precision $\mathcal{P}$ (how correct are the results?), and F-measure $\mathcal{F}$ (harmonic mean of P and R). We perform a ten-fold cross-validation. In the perspective of our work, these measures help evaluate the suitability of the methodology to the distinction between words understandable or not by layman speakers and the relevance of the chosen features to the aimed task. The baseline corresponds to the assignment of words to the biggest category, *e.g.*, *I cannot understand*, which represents 71%. We can also compute the gain, which is the effective improvement of performance P given the baseline BL (Rittman, 2008): $\frac{P-BL}{1-BL}$.

4.2 Processing the Xhosa data

The Xhosa corpus is processed with distributional methods (Harris, 1968; Brown et al., 1992): they aim at grouping words that share the same or similar contexts. As an example, *symptom* and *pain* may be grouped together because they share several common contexts as they appear in the neighborhood of words such as *relieve*, *appear* or *treatment*. It is assumed that such groups of words also have some semantic relations among them, although these relations are not semantically typed. The context is defined as co-occurrence window (n words before and/or after a given word). This context is exploited to compute the association strength between words within the window: it is usually based on frequencies or Mutual Information. A similarity measure (i.e., Jaccard or Cosine) is computed and allows to group words together (Curran, 2004). We apply an implementation of the Brown algorithm[6] with the following parameters:

- corpus content with and without punctuation;
- English text filtered out or not;
- normalization to lower-cased characters or not;
- setting up the minimal number of occurrences of words, within the interval [1, 2, 3];
- setting up the number of clusters to be generated within the interval [50, 100, 150, 200, 250...1000, 1500, 2000, 2500].

[6] http://cs.stanford.edu/~pliang/software/

4.3 Interpreting the Xhosa results

The distributional methods generate clusters which words share common contexts and possibly common semantics. We expect that the clusters we generate with these methods may contain: (1) words belonging to the same syntactic classes (*i.e.*, verbs, nouns); (2) words playing similar syntactic roles (*i.e.*, prepositional phrases); (3) words with similar semantics (*i.e.*, general language words, pathologies, treatments); (4) or even words that represent similar comprehension levels (*i.e.*, easy to understand, difficult to understand). This last possibility, unsupervised automatic distinction between words that represent similar comprehension levels, would make a direct parallel between French and Xhosa data. However, given the difference of the source data (simple contexts and frequencies with Xhosa data, and a set of 24 sophisticated features with French data), we cannot expect to achieve this possibility with the currently exploited methods and resources.

For interpreting the Xhosa results, the generated clusters are annotated using the bilingual lexicon. Two kinds of annotations are performed:

1. *direct:* those cluster words that exist in the lexicon are provided with their English translations;
2. *indirect:* the words that are not recorded in the lexicon are checked for a surface likeness with words from the lexicon. For this, the number of character deletions, insertions and reversals from a given cluster word to obtain a given lexicon word are counted (Levenshtein, 1966). For instance, *ewonke* is not part of the lexicon, but *wonke*, *konke* and *zonke* are. The cost to transform *ewonke* into these lexicon words is 1, 2 and 2, respectively. We have indeed to delete one character *e* to obtain *wonke*, and we have to delete one character *e* and replace one character $w \rightarrow k$ to obtain *konke*. The maximal cost is set to 2, which means that the three candidate translations detected for *ewonke* are acceptable. The meaning of the three candidate translations (*e.g.*, *all*), can be then transposed onto the *ewonke* word.

5 RESULTS AND DISCUSSION

5.1 Relevant features for French

Table 3: Performance and gain obtained for $\mathcal{F}$ by J48.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	BL	gain
Majority	0.876	0.889	0.881	0.71	0.16

Performance of the J48 algorithm is among the best obtained: we use it to present the results. Besides, this algorithm provides the decision tree, which allows analyzing the features exploited for the classification. Performance of J48 on the *majority* dataset is indicated in Table 3: $\mathcal{P}$ 0.876, $\mathcal{R}$ 0.889, and $\mathcal{F}$ 0.881. The gain we obtain is 0.16 point by comparison with the baseline. These are good results and indicate that the chosen features are relevant to the purpose of the study. A more detailed analysis of the influence of the individual features indicate that:

- with the syntactic categories alone we obtain $\mathcal{P}$ and $\mathcal{R}$ between 0.65 and 0.7;
- semantic axes of SNOMED Int decrease precision (be frequent in Snomed Int does not mean to be easier to understand) but improve recall;
- presence of words in the reference lexica is beneficial to both precision and recall;
- the frequencies of the lexemes on the general search engine are often beneficial;
- the suffixes with the three- and four-character length (*i.e.*, *omie, phie, émie*) have a positive impact, but the suffixes with the five-character length negatively impact the results;
- among the features taht negatively impact the results, we find also readability scores (especially the Flesch-Kincaid score) and number and percentage of consonants;
- the remaining features have no or very small impact on the performance.

Notice that features, such as frequency on the search engine or presence in the reference lexica, proved to be efficient in SemEval contest (Specia et al., 2012).

5.2 Generation of clusters for Xhosa

In Table 4, we give quantitative information on clusters generated with the Xhosa data while the number of clusters is set to 1000, 1500, 2000 or 2500, and the minimal frequency of words is set to 2. The total number of words is then 17,890 (punctuation tokenized and removed, English text kept, lower-cased data). We indicate minimal, maximal and average number of words per cluster. We can observe that, logically, the average number of words per cluster is going decreasing with a higher number of clusters. Clusters containing only one word are frequent. Exceptionally, we can also generate clusters with a large number of words (*i.e.*, 115 with 2500 clusters). In the following, we indicate and discuss the results for 1500 clusters. These provide a good compromise of the size of clusters and their content: they are neither too exclusive nor too inclusive. The cluster words are mapped to the lexicon. With 1500 clusters, we obtain 5,873 direct and 12,017 indirect mappings. Among the general observations, we can notice that:

Table 4: Quantitative information on clusters generated with the Xhosa data (minimal frequency set to 2).

Nb clusters	1000	1500	2000	2500
Min/cluster	1	1	1	1
Max/cluster	60	60	54	115
Average	17.89	11.93	8.95	7.16

- when the English sentences are kept, the English words are usually clustered together in separate clusters, which shows the efficiency of unsupervised approaches for language recognition. To give an idea of the English words in the corpus: when all words are taken into account (minimal frequency 1, with upper-cased characters), we find 50,421 words among which 8,401 are in English; with the minimal frequency set to 2, we find 19,405 words among which 3,402 are in English;
- when upper and lower-cased words are used together, the same words written differently (*i.e.*, *iya* and *Iya*) are usually grouped together, which means that their contexts are similar and that their regularity can be observable in both cases.

Concerning more semantic aspects of the clusters, the results indicate that:

- a lot of clusters gather verbal forms, among which those dedicated to verbs with meaning *to go/come* or *to get/give* are very frequent, *i.e.*:
  - *ayiyi (it doesn't go)* and *iya (it goes)*, or *ndiza (I come)*, *sisiya (we go)* and *uza (he comes)*;
  - *bakunike (they gave it to you)*, *amnike (they gave for him)*, *lunike (give it)*, *zinike (give for them)* and *afumane (they got)*;
- cluster words may share common semantics, like:
  - the notion of *improvement*, with words such as *kuyanceda (it helps)*, *kuthenjelwe (you can hope)*; or *kuncede (it helps you)*, *kuxhaswe (you support)*, *unyangwe (he or she was cured)*, *kuthintelwe (prevent)*, *uhlale (he or she lived)*;

  - the notion of *movement inside*, with words such as *kungena (it goes into)*, *ebunzimeni (they are in the mass)*, *zifakwa (they are put in)*;
  - the notion of *interaction*, with words such as *bajongane (look at each other)*, *kuxoxwe (discuss it)*, and *ukuqondwa (understand)*;
- moreover, some clusters contain medical notions, such as *macrophage* and *ziingxaki (they are the problems)*, *zigulane (sick each other)* and *ngozi (danger)*; *kunengozi (it has the danger)* and *mfuneko (it's a requirement/need)*; *agcine (they saved)*, *ndincede (help me)* and *kunyanga (it cures)*. Sometimes, these are related to the *improvement* notion.

The currently obtained results on Xhosa are in accordance with our expectations on their content: we can automatically detect and group together words with general, grammatical and medical notions. The current method cannot distinguish between comprehensible and non-comprehensible words. This means that such distinction cannot rely only on the contexts of words, but requires additional information such as those used in the study on the French data. For this reason, work done on the Xhosa data should be considered as preliminary. Still, the English words are usually clustered separately: typically, such words can represent difficult notions for Xhosa-speaking community (Levin, 2006b; Schlemmer and Mash, 2006). Notice that these results are similar to what we observe on French data: words borrowed from Latin and Greek, or words that are morphologically complex and that contain Latin or Greek bases, are usually felt to be non-understandable by French speakers. Further observation of the difficulty of the Xhosa words will require additional methods and rely on a specific judgment of native speakers.

5.3 Comparison with existing studies

Some findings of the proposed work can be compared with existing studies: relevance of some features used in French (frequency, presence in the reference lexica) to the readability diagnosis (Specia et al., 2012). Still, existing work is usually applied to English data, while processing of French and of less resourced languages (like Xhosa) is rare or even non-existing. Moreover, usually data from the general language are processed and little interest is paid to medical language. For such reasons, it remains difficult to fully compare the proposed study with existing work. We can nevertheless rely on the SemEval contest and on the work done in French to improve the method applied to Xhosa and to design a similar method for this language.

6 CONCLUSIONS AND FUTURE WORK

We proposed experiments in French and Xhosa languages with the main objective to detect comprehensible and non-comprehensible words from the medical field automatically. The material, resources and methods used in both languages are different (non supervised clusters of words in Xhosa, supervised categorization of French words), which logically lead to different results. Nevertheless, we can do similar observations in both languages concerning the detection of borrowed and foreign words (*e.g.*, Latin and Greek words and morphological components in French, English words in Xhosa), which appear to be difficult to understand for native speakers without training in medicine. It appears that the distinction between comprehensible and non-comprehensible words cannot rely only on the contexts of words: additional information and exploitation of supervised methods are necessary. The results obtained for Xhosa are preliminary from this point of view.

Future studies should help explicate these avenues and analyze more complete data. For instance, we plan to study the content of clusters and to address the understandability level of the Xhosa words with the native speakers. We also plan to use more sophisticated clustering approaches and to apply a supervised approach for processing the Xhosa data. For performing this last issue, we need to compute a reasonable set of suitable features, some of which are suggested by the work performed in French and in SemEval contest: frequency of words in the studied corpus or on the web, their frequency in a reference corpus or in a more technical corpus if available, morphological analysis of words (Bosch et al., 2008; Pretorius and Bosch, 2009) and their lexical complexity, complexity of words computed at the character level (*i.e.*, number of characters and syllables, readability scores adapted to the Xhosa language), common contexts of words.

Acknowledgements

Experiments on the French data have been done in part within the MESHS project COMETE.

REFERENCES

Allwood, J., Grönqvist, L., and Hendrikse, A. (2003). Developing a tag set and tagger for the african languages of South Africa with special reference to Xhosa. *Southern African Linguistics and Applied Language Studies*, 21(4):223–237.

AMA (1999). Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, 281(6):552–7.

Berland, G., Elliott, M., Morales, L., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Munoz, J., Puyol, J., Lara, M., Watkins, K., Yang, H., and McGlynn, E. (2001). Health information on the internet. accessibility, quality, and readability in english ans spanish. *JAMA*, 285(20):2612–2621.

Bosch, S., Pretorius, L., and Fleisch, A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, 17(2):66–88.

Brown, P., deSouza, P., Mercer, R., Della Pietra, V., and Lai, J. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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

Curran, J. R. (2004). *From distributional to semantic similarity*. PhD thesis, University of Edinburgh.

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

Dold, A. and Cocks, M. (1999). A preliminary list of Xhosa plant names from the Eastern Cape, South Africa. *Bothalia*, 29:267–292.

Elhadad, N. and Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *BioNLP*.

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

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

Flores, G., Abreu, M., Olivar, M., and Kastner, B. (1998). Access barriers to health care for latino children. *Arch Pediatr Adolesc Med*, 152:1119–1125.

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

Harris, Z. S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Kincaid, J., Fishburne, R. J., Rogers, R., and Chissom, B. (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.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 707(10).

Levin, M. (2006a). Different use of medical terminology and culture-specific models of disease affecting communication between Xhosa-speaking patients and English-speaking doctors at a South African paediatric teaching hospital. *S Afr Med J*, 96:1080–1084.

Levin, M. (2006b). Language as a barrier to care for Xhosa-speaking patients at a South African paediatric teaching hospital. *S Afr Med J*, 96:1076–1079.

Messai, R., Zeng, Q., Mousseau, M., and Simonet, M. (2006). Building a bilingual french-english patient-oriented terminology for breast cancer. In *MedNet*.

Miller, T., Leroy, G., Chatterjee, S., Fan, J., and Thoms, B. (2007). A classifier to evaluate language specificity of medical documents. In *HICSS*, pages 134–140.

Moropa, K. (2007). Analysing the English-Xhosa parallel corpus of technical texts with Paraconc: a case study of term formation processes. *Southern African Linguistics and Applied Language Studies*, 25(1):183–205.

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

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

Pretorius, L. and Bosch, S. (2009). Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *AFLAT*, pages 96–103.

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

Roux, J., Louw, P., and Niesler, T. (2004). The African speech technology project: An assessment. In *LREC*, pages 93–96.

Schlemmer, A. and Mash, B. (2006). The effects of a language barrier in a South Africa district hospital. *S Afr Med J*, 96:1084–1087.

Schmid, H. (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.

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

TLFi (2001). *Trésor de la Langue Française - I*. INaLF/ATILF. Disponible l'adresse `www.tlfi.fr`.

Van der Stouwe, C. (2009). A phonetic and phonological report on the Xhosa language. Technical report. Accessed 1 October 2013, http://bit.ly/1bZwt1j.

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

Woloshin, S., Bickell, N., Schwartz, L., Gany, F., and Welch, H. (1995). Language barriers in medicine in the united states. *JAMA*, 273(9):724–728.

Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J., and Browne, A. C. (2006). Exploring lexical forms: first-generation consumer health vocabularies. In *AMIA 2006*, pages 1155–1155.

Zeng, X. and Parmanto, B. (2003). Evaluation of web accessibility of consumer health information websites. In *AMIA 2003*, pages 743–7.