Automatic Detection of Difficulty of French Medical Sequences in Context

Anaïs Koptient, Natalia Grabar
CNRS, Univ Lille, UMR 8163 - STL
F-59000 Lille, France
{anais.koptient, natalia.grabar}@univ-lille.fr

Abstract

Medical documents use technical terms (single or multi-word expressions) with very specific semantics. Patients may find it difficult to understand these terms, which may lower their understanding of medical information. Before the simplification step of such terms, it is important to detect difficult to understand syntactic groups in medical documents as they may correspond to or contain technical terms. We address this question through categorization: we have to predict difficult to understand syntactic groups within syntactically analyzed medical documents. We use different models for this task: one built with only internal features (linguistic features), one built with only external features (contextual features), and one built with both sets of features. Our results show an f-measure over 0.8. Use of contextual (external) features and of annotations from all annotators impact the results positively. Ablation tests indicate that frequencies in large corpora and lexicon are relevant for this task.

Keywords: Syntactic Groups, Complexity Detection, Linguistic and Contextual Features, Medical, French

1. Introduction

As any specialized area, medical domain witnesses different types of actors, all involved in the healthcare process and biomedical research, such as medical doctors, patients, nurses, biologists, medical students, or pharmacists. Patients particularly have no particular medical knowledge and may have understanding problems when reading medical information. Indeed, medical domain uses technical terms, such as *cholestatic jaundice* or *mesenteric venous thrombosis*. Such terms have specific and opaque semantics. Yet, the understanding of these notions is crucial for patients as it is intimately linked to their healthcare and wellbeing. It has indeed been shown that a correct understanding of medical notions plays an important role in healthcare process and ensures its success (Hermann et al., 1978; Vander Stichele, 2004; Mcgray, 2005; Eysenbach, 2007). It has also been shown that patients have to face quite frequently technical medical documents, in which the level of technicality is above their understanding:

- information on drug intake, preparation and dosage (Vander Stichele, 1999; Patel et al., 2002);
- clinical documents (Vander Stichele, 1999; Patel et al., 2002) on clinical procedures;
- medical leaflets and consent forms (Williams et al., 1995), specifically created for and typically met by patients during their healthcare process;
- more generally, information for patients found on the Internet (Rudd et al., 1999; Berland et al., 2001; Mcgray, 2005; Oregon Practice Center, 2008; D'Alessandro et al., 2001; Brigo et al., 2015) on different medical topics.

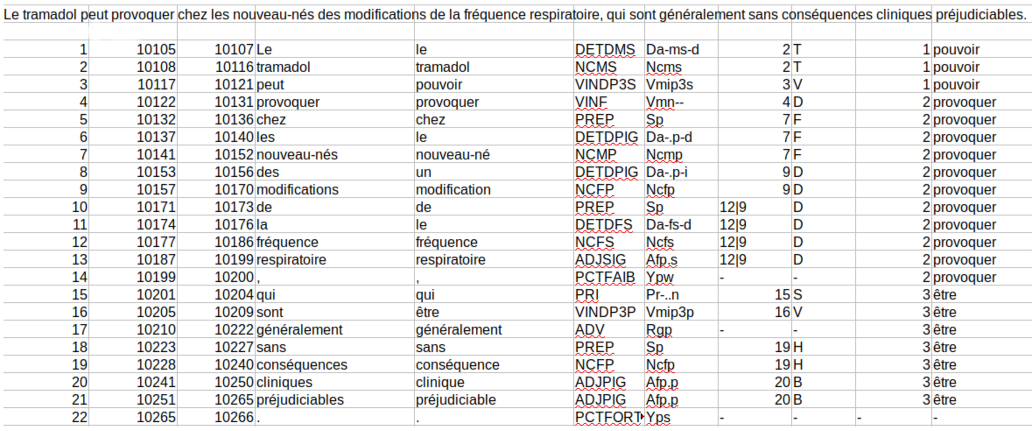
Thus, it is important to detect terms and syntactic groups that can show understanding difficulties for patients. Those terms can then be simplified. In this work, we propose a contribution to this research question: identification of difficult to understand syntactic groups in French medical texts. We first introduce existing works on this question (section 2). We then present the material used (section 3). Next, we describe the method proposed (section 4). Finally, we present the results (section 5) and discuss them (section 6).

2. Related work

Several works have been done throughout the years on the prediction of the difficulty in whole documents (Zheng et al., 2002; Chmielik and Grabar, 2009; Vajjala and Meurers, 2015) and they show good scores, with F-measures higher than 0.9 when different features are used. Indeed, at the text level, several hints are available and give complementary results. Nevertheless, prediction of difficulty of terms and syntactic groups within sentences is a more complex issue.

Works on this issue mainly use supervised learning classifiers with features including linguistic (frequency, length of the word, part-of-speech, number of phonemes, of syllables, phoneme/spelling coherence...) and psycholinguistic (level of abstractness) features (Paetzold and Specia, 2016; Yimam et al., 2018; Gala et al., 2013; Shardlow, 2013; Sheang, 2019; Agarwal and Chatterjee, 2021), as well as word embeddings and contextual features (Yimam et al., 2018; Sheang, 2019). Other works focus on exploitation of frequency. In particular, frequency thresholding is important (Zeng et al., 2005), as the frequency of words is considered to be a good hint to determine their complexity (Leroy et al., 2013; Lindqvist et al., 2013; Rudell, 1993). Another work suggests that the rarity of words may be indicative about their difficulty: the words that are not found in different lexica are considered to be difficult (Borst et al., 2008). (Zaharia et al., 2020) proposed a method using RNN and Transformer-based models. Finally, more recent works use Bert models (Shardlow et al., 2021).

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 55–66
Marseille, 25 June 2022.



Le tramadol peut provoquer chez les nouveau-nés des modifications de la fréquence respiratoire, qui sont généralement sans conséquences cliniques préjudiciables.

1	10105	10107	Le	le	DETDMS	Da-ms-d	2	T	1	pouvoir
2	10108	10116	tramadol	tramadol	NCMS	Ncms	2	T	1	pouvoir
3	10117	10121	peut	pouvoir	VINDP3S	Vmip3s	3	V	1	pouvoir
4	10122	10131	provoquer	provoquer	VINF	Vmn--	4	D	2	provoquer
5	10132	10136	chez	chez	PREP	Sp	7	F	2	provoquer
6	10137	10140	les	le	DETDPIG	Da-.p-d	7	F	2	provoquer
7	10141	10152	nouveau-nés	nouveau-né	NCMP	Ncmp	7	F	2	provoquer
8	10153	10156	des	un	DETDPIG	Da-.p-i	9	D	2	provoquer
9	10157	10170	modifications	modification	NCFP	Ncfp	9	D	2	provoquer
10	10171	10173	de	de	PREP	Sp	12|9	D	2	provoquer
11	10174	10176	la	le	DETDFS	Da-fs-d	12|9	D	2	provoquer
12	10177	10186	fréquence	fréquence	NCFS	Ncfs	12|9	D	2	provoquer
13	10187	10199	respiratoire	respiratoire	ADJSIG	Afp.s	12|9	D	2	provoquer
14	10199	10200	,	,	PCTFAIB	Ypw	-	-	2	provoquer
15	10201	10204	qui	qui	PRI	Pr-..n	15	S	3	être
16	10205	10209	sont	être	VINDP3P	Vmip3p	16	V	3	être
17	10210	10222	généralement	généralement	ADV	Rgp	-	-	3	être
18	10223	10227	sans	sans	PREP	Sp	19	H	3	être
19	10228	10240	conséquences	conséquence	NCFP	Ncfp	19	H	3	être
20	10241	10250	cliniques	clinique	ADJPIG	Afp.p	20	B	3	être
21	10251	10265	préjudiciables	préjudiciable	ADJPIG	Afp.p	20	B	3	être
22	10265	10266	.	.	PCTFORT	Yps	-	-	-	-

Figure 1: Syntactic annotation and parsing from Cordial

The main contributions of our work are:

- building annotations of understanding difficulties in French medical documents,
- automatic prediction of understanding difficulties in French medical documents,
- exploitation of internal (linguistic) and external (contextual) features,
- study of the impact when using annotations from several annotators.

3. Material

We use 100 French clinical cases randomly selected from the CAS corpus (Grabar et al., 2018), including a total of 41,384 words. Clinical cases are medical documents similar to clinical reports. They describe the patients medical background, the reason of their consultation, healthcare process and treatments proposed and performed, and the outcome. Such clinical documents can be encountered by patients in their everyday lives. Clinical cases deal with different topics and specialties. They are published and are freely accessible in different sources. They are anonymous.

The corpus with clinical cases is pre-processed. The documents are syntactically analyzed by Cordial parser (Laurent et al., 2009) to divide them into syntactic groups. Figure 1 shows the output from Cordial. We exploit the following syntactic information: the first column with the id of the word within the sentence, and the eighth column with the id of the head of the syntactic group in which the word belongs (words with the same number belong to the same syntactic group). For instance, {*Le tramadol*; *the tramadol*} is a syntactic group where *tramadol* is the head. When a given word belongs to a group within a group, we keep the minimal one, that is, the group within the bigger group. The corpus provides in total 15,053 syntactic groups. The choice to work with syntactic groups instead of words is motivated by the fact that syntactic groups may cover single or multi-word expressions, which convey specific semantics (Baldwin and Kim, 2010) and represent then suitable processing units.

Documents are then annotated manually by nine annotators. The annotators are all native French speakers. They have no medical knowledge or training. Few of them (annotators 5 to 8) are chronically ill with hemophilia, while others have no chronic disorders. The annotators were advised not to use dictionaries or Internet when annotating. They had to do the annotations on the basis of their own knowledge. The annotators are presented with whole documents, where syntactic groups are between brackets, such as indicated on Figure 2. For each syntactic group, the annotators have to indicate if they do not understand it (by annotating it as *not-understood*) or if they are not sure to understand it (by annotating it as *not-sure-to-understand*). In the case they understand a given syntactic group, they do not have to annotate it.

[Her medical background] [shows] [a probable gestational diabetes] [and a HG] [during her first pregnancy]. [The patient] [had then been hospitalized] and [recieved] [an intravenous treatment] [of metoclopramide with] [diphenhydramine followed] [by oral treatment] [with metoclopramide and] [hydroxyzine]. [An extrapyramidal reaction] ([jaw] [stiffness and] [difficulty] [to talk]) [caused] [the cessation] [of metoclopramide]. [Hydroxyzine] [had] [then been replaced] [by the combination] [of doxylamine] [and pyridoxine] (Di-clectinMD).

Figure 2: Translated excerpt from syntactically parsed and annotated clinical case

Further to the annotation process, each document is annotated by at least four annotators, while some documents are annotated by up to six annotators. We computed the kappa of Fleiss (Fleiss, 1971) for four annotators who annotated all the documents. As indicated in Table 1, the kappa for all annotators is 0.175, which is a low value. For some pairs of annotators (1&3, 2&3), kappa shows slightly higher values (0.292 and 0.316).

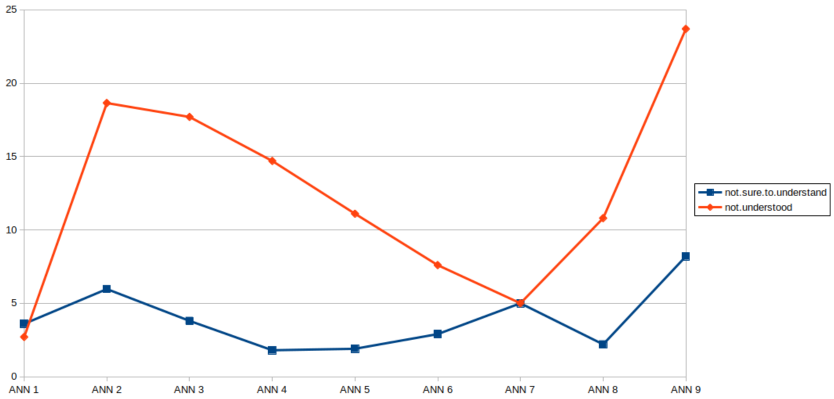


Figure 3: Percentages of *not sure to understand* (blue line) and *not understood* (red line) annotations according to the annotators. Annotators 5 to 8 are chronically ill

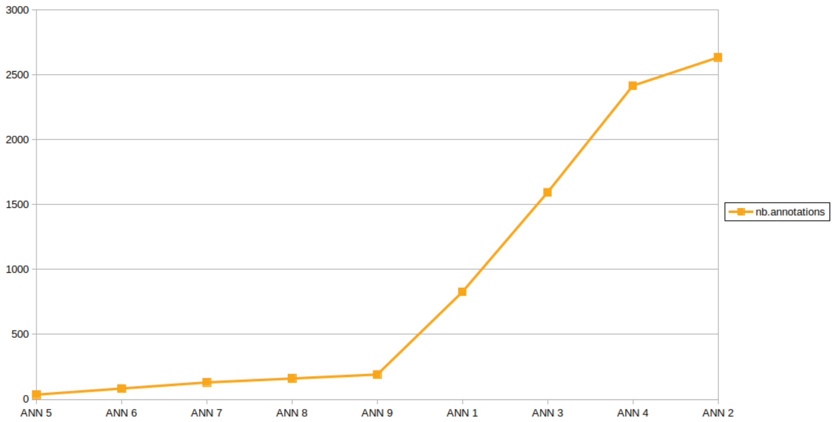


Figure 4: Number of different annotations (*not sure to understand* and *not understood*) from each annotator

We assume this means that the task at hand is very subjective. Besides, it is impossible to do the consensus among the annotators and to convince them that they should understand a given syntactic group. Indeed, this kind of annotations heavily depends on own knowledge and understanding feeling of each person.

Annotators	*Kappa*
all (1-4)	0.175
1 & 2	0.093
1 & 3	0.292
1 & 4	0.1
2 & 3	0.316
2 & 4	0.115
3 & 4	0.048

Table 1: Kappa score for different annotators

Figure 3 shows the percentage, for each annotator, of *not sure to understand* (blue line) and *not understood* (red line) annotations. We can see that annotators who are chronically ill (annotators from 5 to 8) have a lower percentage of *not sure to understand* and *not understood* annotations. For instance, Annotator 7 marked only 5% of syntactic groups as *not understood* and that much syntactic groups as *not sure to understand*. We assume that chronically ill annotators may better understand medical terms than healthy annotators.

Another interesting observation is that the annotations are complementary. Hence, Figure 4 shows the number of new annotations (*not sure to understand* and *not understood*) from each annotator, starting with chronically ill annotators who annotated the lowest number of non-understandable syntactic groups. We can see that the number of different and new annotations is increasing as a new annotator is taken into account. As noticed above, the feeling on understanding difficulty of medical information is a subjective question which depends on the own knowledge and individual experience of annotators. We consider that, in order to obtain a more complete picture on understanding difficulties, it would be necessary to involve a greater number of annotators: they may contribute with more relevant annotations for a given population. In this case, the purpose is not to achieve a better inter-annotator agreement but to obtain the more complete annotations possible.

When the annotations are done, we merge them all together. For this, we keep the strongest annotation for a given syntactic group: if one annotator annotates a given syntactic group as *not understood*, while all the others annotate it as *understood*, we therefore consider this syntactic group as *not understood*. In total, 12,417 syntactic groups belong to the *understood* category, 157 belong to the *not sure to understand* category, and 2,479 belong to the *not understood* category. We decide to merge together *not sure to understand* and *not understood* categories because: the *not sure to understand* category is very small and the difference between these two categories lays in the certainty related to the non-understanding of syntactic groups. This disposition permits also to do a binary classification task.

Figure 2 presents an English translation from annotated clinical case. Syntactic groups are between brackets. Groups in red are annotated as *not understood*, and groups in blue as *not sure to understand*. Hence, we obtain a French dataset with 15,053 syntactic groups annotated according to their difficulty. This dataset is divided into training (75%) and test (25%) sets.

4. Determining the difficulty of syntactic groups in context

We address the prediction of difficulty of syntactic groups as categorization problem: for a given syntactic group, we have to decide if it should be assigned to the category *not understood* or to the category *understood*. We first introduce our approach for determining the difficulty of syntactic groups in context and then describe the experimental setup.

4.1. Approach

We test several supervised learning algorithms implemented in Scikit-Learn (Pedregosa et al., 2011) to determine the difficulty of French medical syntactic groups in context: SVM Linear and RBF (Platt, 1998), Decision Tree (Quinlan, 1993), Multilayer Perceptron (Rosenblatt, 1958), and Random Forest (Breiman, 2001). These classifiers have been used for similar tasks in previous works (Ronzano et al., 2016; Mukherjee et al., 2016; Zampieri et al., 2016; Brooke et al., 2016; Davoodi and Kosseim, 2017; Alfter and Pilán, 2018; Kajiwara and Komachi, 2018) and display accuracies between 0.513 and 0.933.

We exploit internal and external features. Internal features are related to internal and linguistic properties of syntactic groups:

- *Number of letters*. Previous studies have shown that word length correlates with simplicity of text (Keskisärkkä, 2012). Moreover, simplification guidelines (Ruel et al., 2011; OCDE, 2015; UNAPEI, 2019) preconize to use short terms;
- *Number of phonemes*. Number of phonemes is correlated with word length. To determine the number of phonemes, we used the French database Lexique3 (New et al., 2001) and the French adaptation of the Epitran Python module (Mortensen et al., 2018);
- *Number of syllables*, which is, once again, correlated with word length. To determine it, we also use Lexique3 and Epitran;
- *Coherence between spelling and number of phonemes*. This feature corresponds to the ratio between the number of phonemes and the number of letters. Its values are between 0 and 2. If there is no difference then the coherence value is 0, if there is one or two differences the coherence value is 1, and if there are more than two differences the coherence value is 2;
- *Syllable components*. This feature corresponds to three levels of complexity according to the syllable components (coined with consonants C, vowels V and semi-consonants Y) and to their frequency. For instance, syllables like CYV *lion* (lion), CVC *mentir* (to lie), CV *lettre* (letter) are very frequent in French, while syllables like CCVC *attendrir* (to soften), VCC *ans* (years), VC *antan* (yesteryear), YV *ion* (ion) are much less frequent in French;
- *Frequency*. Several studies show that the complexity of words can be related to their frequency (Leroy et al., 2013; Lindqvist et al., 2013; Rudell, 1993). We use several sources to compute the frequency:
  - frequency in French lexica: Lexique3 and Manulex (Lété et al., 2004),
  - frequency in a general language corpus (French Wikipedia),
  - frequency in a medical corpus (CLEAR corpus (Grabar and Cardon, 2018)).

  For syntactic groups containing more than one word, we compute the average of frequencies of each word.
- *Presence of words in a list of very basic French vocabulary* built by Catach (Catach, 1984).

Notice that several of these features are inspired by a typology in a related work (Gala et al., 2013).

Among the external features, we count the right and left contexts of the syntactic groups. Hence, for each syntactic group, we extract five words at its left and five at its right, within the sentence.

We build a bi-class model, where each class comes from the manual annotations: *not understood* corresponds to *not understood* and *not sure to understand*; and *understood* corresponds to *understood*.

4.2. Baseline

For the baseline approach, we exploit the UMLS (Unified Medical Language System) (Lindberg et al., 1993):

- if a given syntactic group is present in the UMLS this group is considered as *not understood*. Indeed, in this case, the syntactic group is part of the specialized terminology and may be considered to convey technical meaning,
- if a given syntactic group is not present in the UMLS it is considered as *understood*. In this case, the syntactic group may be considered to convey more general meaning.

4.3. Experimentations

We use supervised learning algorithms with: only internal features, only external features, both internal and external features. We also perform ablation tests: (1) only one feature is used and the remaining features are removed, (2) one feature is removed.

Each experimentation is evaluated within the training dataset through 10-fold cross-validation using recall, precision and f-mesure. Since the classes are unbalanced in the training set (1,978 instances in the *not-understood* class and 10,294 instances in the *understood* class), we train other models on a balanced training set (1,978 instances in *not-understood* and *understood* classes). The 1,978 *understood* instances are selected randomly within the 10,294 *understood* instances from the full train set. In addition, the models built on both training sets (full set and the one with balanced classes) are tested on the test set, and recall, precision and f-mesure are also computed. All results are compared to the baseline.

Besides, all features are exploited with annotations from each annotator used incrementally. The purpose is to observe a possible impact on categorization results when using more annotators.

5. Results

Among the classifiers tested, Random Forest provides the best results in several settings. Also, contrary to other classifiers, it tries to recognize the two categories (*not understood* and *understood*) and not only the largest category (*understood*). Hence, we present the results obtained with this classifier. We first present the classification results obtained with ten-fold cross-validation and on the test set (Section 5.1), we then describe the results of the ablation tests (Section 5.2).

5.1. Classification of syntactic groups

Table 2 shows the results of the ten-fold cross-validation on balanced training set depending on the features used (internal, external, or both) and compared to the baseline. The baseline scores are very low, and this can be explained by the fact that any word linked to medical domain is present in the UMLS, even those that can be understood by non-medical experts. For instance, {*anestésie*; *anesthesia*} is annotated as *understood* in the reference data but is considered as *not-understood* by the baseline method because this term is part of the UMLS. All feature sets outperform the baseline. More specifically, the combination of both sets of features provides the highest scores (0.931 precision, 0.847 recall and 0.877 f-measure) in this setting. With the three sets of features, the values of precision and recall are close to each other.

Model	*Precision*	*Recall*	*F-measure*
Internal	0.805	0.769	0.783
External	0.893	0.798	0.830
Both	0.931	0.847	0.877
Baseline	0.570	0.579	0.573

Table 2: Results of the ten-fold cross-validation with different feature sets and Random Forest obtained on the full training set

Table 3 shows the results obtained on the test set with different models trained on the full training set: internal and external features, both of them, and the baseline. The combination of both external and internal features gives once again the higher scores. Yet, for all models, the scores become lower, and the baseline outperforms other models.

Model	*Precision*	*Recall*	*F-measure*
Internal	0.384	0.500	0.434
External	0.598	0.524	0.248
Both	0.601	0.551	0.310
Baseline	0.567	0,570	0.567

Table 3: Evaluation on the test dataset with different feature sets and Random Forest on the full training set

Table 4 shows the results of the ten-fold cross-validation obtained on the balanced training set with different features used (internal, external, or both) and compared to the baseline. The scores are lower than those obtained on the full training set (see Table 2). The combination of internal and external features outperforms other feature sets. All models outperform the baseline.

Model	*Precision*	*Recall*	*F-measure*
Internal	0.734	0.734	0.734
External	0.730	0.703	0.707
Both	0.798	0.799	0.798
Baseline	0.570	0.579	0.573

Table 4: Results of the ten-fold cross-validation with different feature sets and Random Forest on the balanced training set (both classes are equivalent)

Table 5 shows the results obtained on the test set with different models trained on the balanced training set with different feature sets (internal and external features, both of them), and the baseline. The scores are lower than those obtained on the full training set (see Table 3). The baseline outperforms other models.

Model	*Precision*	*Recall*	*F-measure*
Internal	0.602	0.505	0.101
External	0.384	0.500	0.434
Both	0.407	0.470	0.428
Baseline	0.567	0,570	0.567

Table 5: Evaluation on the test dataset with different feature sets and Random Forest on the balanced training set (both classes are equivalent)

5.2. Ablation tests

We performed two ablation tests: (1) only one feature is exploited and the remaining features are removed, and (2) one feature is removed at a time from the whole feature set. These ablation tests are done with internal features and are evaluated by a ten-fold cross-validation. We compare these results with the baseline and exploitation of all internal features.

Figure 5 shows f-measure when only one feature is used (burgundy line). The features indicated on the horizontal axis are the features which are kept. We compare these results to the exploitation of all internal features (green line) and baseline (yellow line). As already observed, the baseline outperforms the use of internal features only. We can also see that combination of all internal features (green line) is more efficient than each feature taken alone. We observe that the scores become lower with several features used individually: cohesion feature, number of letters and number of syllables, the Catach list, and syllable components. We can provide an explanation on these observations:

- the length of words and syntactic groups is not always correlated with their complexity in medical documents, contrary to long words from the general language texts. Indeed, short medical words, like abbreviations or some medical terms, can correspond to complex notions, while long words do not necessarily correspond to complex terms;
- the Catach list is very short and covers only a small portion of words occurring within medical documents, contrary to lists from Lexique3 and from Wikipedia which are more exhaustive;
- information on syllables (their structure and cohesion) has been first proposed for the classification of scholar manuals from elementary school, in which this information is important and reflects the scholar levels. We assume, these features are less efficient when used on specialized contents: the overall structure of words and syllables becomes more complex when addressing adult population and is no more a salient feature.

Several features related to the frequency of words provide high scores when used individually: frequency in Lexique3, Manulex, in a general language (French Wikipedia) and medical (CLEAR) corpora. This may be due to the fact that (1) these corpora provide a better coverage for words occurring in medical documents, and (2) the words that have higher frequency in these corpora are also more frequent in the language. Hence, they are better understood by the annotators.

Figure 6 shows f-measure obtained when one feature is removed (burgundy line). The features on the horizontal axis are those features which are removed in a given ablation test. We also present the f-measure when all internal features are used (green line) and the baseline (yellow line). Overall, we can see that the scores become lower when one feature is removed, which indicates that each feature is contributing to the results and that their combination is important. Among the features which removal decreases the scores we can find: the frequency in Lexique3, the number of letters, the frequency on Wikipedia and CLEAR corpora, the syllable components. The impact of the frequency from large corpora (Wikipedia, CLEAR, Lexique3) has already been observed and remains coherent with our observations above. The impact of the number of letters and syllable structure is not observed when these features are exploited individually. Yet, they may find their importance in combination with other features.

Figure 7 shows precision, recall and f-measure from ten-fold cross-validation with incremental addition of annotations from each annotators. Globally, with more annotators the scores progressively become better despite the low inter-annotator agreement. We assume that this group of annotators provides annotations which are complementary and which remain coherent.

6. Discussion

We present an error analysis, and discuss the ablation tests performed. We also compare our work with previously published results.

6.1. Error analysis

We randomly selected eight terms, single words (*furosémide* (furosemide), *sevrage* (withdrawal), *hospitalisée* (hospitalized), *ascite* (ascites)) and multi-word expressions (*chlorure chlorobutanol* (chlorobutanol chloride), *oppression thoracique* (chest tightness), *méga-uretère* (mega-ureter), *pré-opératoire* (preoperative)), to analyze the predictions for these terms. Hence, Table 6 shows the reference annotations, and the predictions provided by the baseline and the models based on internal, external and all the features.

- With internal features, either on full or balanced train set, the syntactic groups are classified as *not understood*, which is the minority category. The model trained on the full training set puts 3,424 out of 3,739 syntactic groups in the *not understood* class. The model trained on the balanced training set puts 3,707 out of 3,739 syntactic groups in the *not understood* class. Therefore, the model trained on the full training set seems to

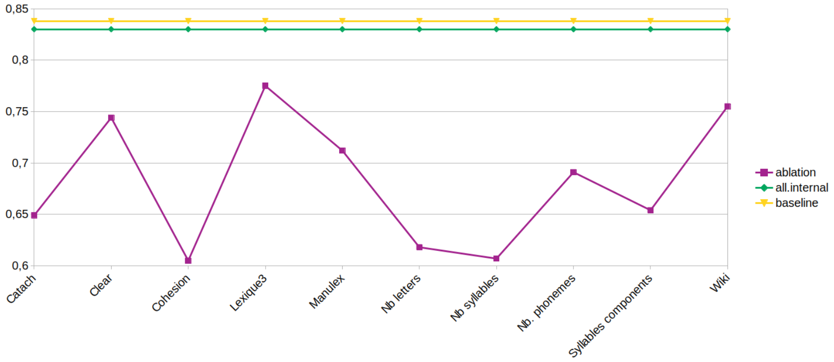


Figure 5: F-measure when only one features is exploited at a time

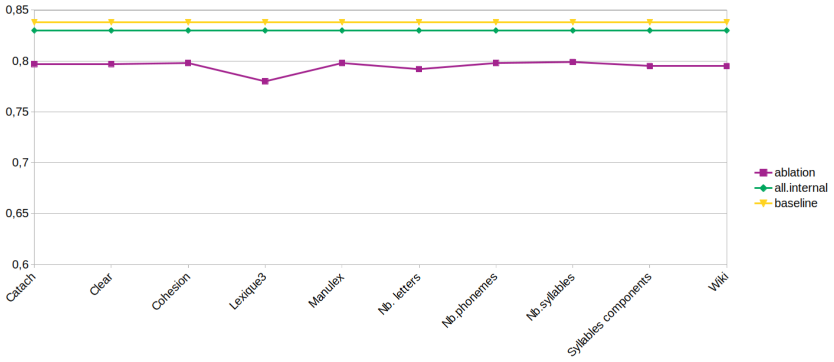


Figure 6: F-measure when one feature is removed

perform better.

- The model based on external features trained on full training set provides wrong predictions for *chlorure chlorobutanol* (chlorobutanol chloride) surprisingly classified as *understood*, and *pré-opératoire* (preoperative) classified as *not understood* certainly because of its length. But overall, this model shows a good performance. The model trained on balanced set classified every syntactic group as *understood*, a non-majority class.

- The model which exploits all the features and is trained on full training set classifies all single-word syntactic groups correctly excepting *hospitalisée* (hospitalized) classified as *not understood* probably because of its length. However, multi-word expressions are all classified as *not understood*. This classification error may also be due to their length. The model trained on balanced training set classified the majority of the syntactic groups as *not-understood* (3,579 out of 3,740).

We assume that low scores obtained when using balanced training set is due to the fact that it contains lower number of instances. However, we believe that the scores can be higher with a larger balanced training set. The baseline only depends on the presence of terms within the UMLS and their recognition. Per se, this is not a very reliable clue because the UMLS is very inclusive. For instance, *sevrage* (withdrawal), which is part of the UMLS, is wrongly predicted as *not understood*. Besides, we also observed that multi-word expressions present a greater challenge for the classification models. Typically, their length may become a confusing feature.

6.2. Ablation tests

According to the ablation tests, frequencies in large corpora (Wikipedia and CLEAR corpora) and lexica (Lexique3) appear to be important features: when removed f-measure decreases while their individual exploitation provides competitive results. As we observed, the size of corpora and lexica may be important as this guarantees that a higher number of words is represented. Besides, their contents may also be important. For instance, the frequencies in Lexique3 are compiled from movie and tv-show subtitles as well as from a book corpus (New et al., 2001), while the frequencies in Manulex are compiled from French scholar books from different levels in primary school. Since Manulex aims to describe children literacy and reading capacity, its exploitation for the analysis of documents written for adults is less useful. The importance of the frequency for the recognition of difficult to understand words has been noticed by several existing works. Indeed, existing work stresses on importance of this feature (Zeng et al., 2005), while several other works ex-

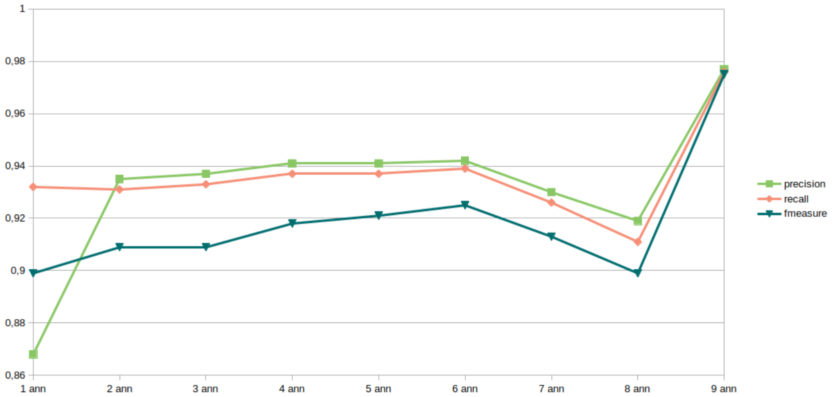


Figure 7: Evaluation measures with incremental addition of annotators, ten-fold cross-validation

syntactic group	*Ref.*	*BL*	*Int. full*	*Ext. full*	*Both full*	*Int. balanced*	*Ext. balanced*	*Both balanced*
furosémide (furosemide)	NU	NU	NU	NU	NU	NU	U	NU
sevrage (withdrawal)	U	NU	NU	U	U	NU	U	NU
hospitalisée (hospitalized)	U	U	NU	U	NU	NU	U	NU
ascite (ascites)	NU	NU	NU	NU	NU	NU	U	NU
chlorure chlorobutanol (chlorobutanol chloride)	NU	NU	NU	U	NU	NU	U	NU
oppression thoracique (chest tightness)	U	NU	NU	U	NU	NU	U	NU
méga-uretère (mega-ureter)	NU	U	NU	NU	NU	NU	U	NU
pré-opératoire (preoperative)	U	U	NU	NU	NU	NU	U	NU

Table 6: Predictions for some syntactic groups (NU: not understood, U: understood)

Previous work	*Feature(s) in common*	*Evaluation*	*F-measure*
(Zampieri et al., 2016)	number of letters	test corpus	0.270
(Ronzano et al., 2016)	number of letters, frequencies	cross-validation	0.735-0.824
(Alfter and Pilán, 2018)	number of letters, number of syllables, frequencies	cross-validation	0.726-0.862
(Alfter and Pilán, 2018)	number of letters, number of syllables and frequencies	test corpus	0.627-0.833
(Kajiwara and Komachi, 2018)	number of letters and frequencies	test corpus	0.745-0.863
(Brooke et al., 2016)	frequencies	test corpus	0.335
(Mukherjee et al., 2016)	number of syllables and presence in basic vocabulary list	test corpus	0.250
(Mukherjee et al., 2016)	number of syllables and presence in basic vocabulary list	cross-validation	0.530
Our work	internal features	test corpus	0.434
	external features	test corpus	0.248
	both	test corpus	0.310
on full training set	internal features	cross-validation	0.783
	external features	cross-validation	0.830
	both	cross-validation	0.877
	internal features	test corpus	0.101
	external features	test corpus	0.434
	both	test corpus	0.428
on balanced training set	internal features	cross-validation	0.734
	external features	cross-validation	0.707
	both	cross-validation	0.798
Baseline	UMLS	test corpus	0.567

Table 7: Comparison with previous works

ploited the frequency for the categorization task (Bingel and Bjerva, 2018; Bingel et al., 2016; Malmasi et al., 2016; Alfter and Pilán, 2018; Kajiwara and Komachi, 2018; Brooke et al., 2016). Besides, one work in French also exploits the frequency from Lexique3, and notices that this feature is important for the task (Gala et al., 2013).

Another observation from the ablation tests is that the number of letters and syllables is less important, although previous works indicate their importance (Gala et al., 2013; Wani et al., 2018). We observe that, even if some features seem to be less important than others individually, both ablation tests indicate that the combination of features improves the results.

6.3. Comparison with previous works

Table 7 shows a comparison with previous similar works, all done with data in English. We consider here the works that have at least one feature in common with our approach. We indicate whether the evaluation is done on a testset or by cross-validation. The comparison is done in terms of the f-measure values. Our results obtained with cross-validation on the full training set are competitive: they are usually higher than those from other works. Results obtained with cross-validation on the balanced training set are closer to those from other works. Finaly, our predictions on test corpus are less competitive yet they overpass several existing works.

7. Conclusion

We proposed to detect difficult syntactic groups in French medical texts thanks to their context (external features) and to their lexical properties (internal features). We use supervised learning algorithms, among which Random Forest appeared to be the best classifier for the task. The models are trained on clinical cases manually annotated according to the difficulty to understand syntactic groups. The dataset is divided in two datasets: training (75%) and test (25%) datasets. We perform several experiments on both full and balanced training sets: exploitation of only internal features (number of letters, number of phonemes, frequencies in corpora and lexica, etc.), exploitation of only external features (five word context on left and right), and of both sets of features. Our baseline is based on the UMLS: if a given syntactic group is part of the UMLS then it is considered as not understood, otherwise it is considered as understood. Two evaluations are performed: ten-fold cross-validation and evaluation on the test dataset. These two evaluations are compared to the baseline. Cross-validation tests indicate that the models built with two sets of features are the most efficient for the task. They shows up to 0.903 f-measure when trained on the full training set and 0.798 f-measure when trained on the balanced training set. However, when all features are exploited on the test dataset, they give relatively low results (0.310 f-measure for the model built on the full training set and 0.428 f-measure on the model built on the balanced training set). We also notice that the reference annotations show low inter-annotator agreement, instead they are complementary: the use of annotations from all annotators progressively improves classification results.

We performed two ablation tests, one where only one feature is kept, and one where one feature is removed at a time. Results of these tests show that the frequency in large corpora and lexica is important, and that word length and number of syllables are less important. We assume that these features require to be combined with other features to show their positive impact on the results. The ablation tests also showed that all features are important, because the best f-measure is obtained when all features are present. We also observed that multi-word expressions present a greater challenge for the classification models. Typically, their length may become a confusing classification feature.

In future work, we plan to enrich the reference dataset with more annotations. As observed, additional annotators enrich the annotated syntactic groups, which improves the classification results. A larger set with the reference data will permit to use approaches involving the Transformers. Besides, as similar datasets are available in other languages (Shardlow et al., 2021; Yimam et al., 2018), we may test our approach on these datasets. Another possible improvement is related to a better consideration of multi-word expressions.

8. Acknowledgements

This work was funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01.

9. Bibliographical References

Agarwal, R. and Chatterjee, N. (2021). Gradient boosted trees for identification of complex words in context. 09.

Alfter, D. and Pilán, I. (2018). SB@GU at the complex word identification 2018 shared task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 315–321, New Orleans, Louisiana, June. Association for Computational Linguistics.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Berland, G., Elliott, M., Morales, L., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Munoz, J., Puyol, J., and et al, M. L. (2001). Health information on the Internet. Accessibility, quality, and readability in English ans Spanish. *JAMA*, 285(20):2612–2621.

Bingel, J. and Bjerva, J. (2018). Cross-lingual complex word identification with multitask learning. In

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 166–174, New Orleans, Louisiana, June. Association for Computational Linguistics.

Bingel, J., Schluter, N., and Martínez Alonso, H. (2016). CoastalCPH at SemEval-2016 task 11: The importance of designing your neural networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1028–1033, San Diego, California, June. Association for Computational Linguistics.

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

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brigo, F., Otte, M., Igwe, S., Tezzon, F., and Nardone, R. (2015). Clearly written, easily comprehended ? The readability of websites providing information on epilepsy. *Epilepsy & Behavior*, 44:35–39.

Brooke, J., Uitdenbogerd, A., and Baldwin, T. (2016). Melbourne at SemEval 2016 task 11: Classifying type-level word complexity using random forests with corpus and word list features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 975–981, San Diego, California, June. Association for Computational Linguistics.

Catach, N. (1984). *Liste Orthographique de Base*. Éditions Nathan, Paris.

Chmielik, J. and Grabar, N. (2009). Comparative study between expert and non-expert biomedical writings: their morphology and semantics. *Stud Health Technol Inform.*, 150:359–63.

D'Alessandro, D., Kingsley, P., and Johnson-West, J. (2001). The readability of pediatric patient education materials on the world wide web. *Arch Pediatr Adolesc Med.*, 155(7):807–12.

Davoodi, E. and Kosseim, L. (2017). Clac at semeval-2016 task 11: Exploring linguistic and psycholinguistic features for complex word identification. *CoRR*, abs/1709.02843.

Eysenbach, G. (2007). Poverty, human development, and the role of eHealth. *J Med Internet Res*, 9(4):34–4.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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

Grabar, N. and Cardon, R. (2018). Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11.

Grabar, N., Claveau, V., and Dalloux, C. (2018). Cas: French corpus with clinical cases. In *LOUHI 2018*, pages 1–12, Bruxelles, Belgique.

Hermann, F., Herxheimer, A., and Lionel, N. (1978). Package inserts for prescribed medicines: what minimum information do patients need? *Br Med J*, 2(6145):1132–1135.

Kajiwara, T. and Komachi, M. (2018). Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana, June. Association for Computational Linguistics.

Keskisärkkä, R. (2012). *Automatic Text Simplification via Synonym Replacement*. Master thesis, Linköping University, Linköping, Sweden.

Laurent, D., Nègre, S., and Séguéla, P. (2009). L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.

Leroy, G., Kauchak, D., and Mouradi, O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform*, 82(8):717–730.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments and Computers*, 36:156–166.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4):281–291.

Lindqvist, C., Gudmundson, A., and Bardel, C. (2013). A new approach to measuring lexical sophistication in l2 oral production. *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, pages 109–126, 01.

Malmasi, S., Dras, M., and Zampieri, M. (2016). LTG at SemEval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000, San Diego, California, June. Association for Computational Linguistics.

Mcgray, A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Mukherjee, N., Patra, B. G., Das, D., and Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 task 11: Identifying complex words in a sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 986–990, San Diego, California, June. Association for Computational Linguistics.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : Lexique//a lexical database for contemporary french : Lexique. *Annee Psychologique - ANNEE PSYCHOL*, 101:447–462, 01.

OCDE. (2015). *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing.

Oregon Practice Center. (2008). Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved. Technical report, Agency for healthcare research and quality. Oregon Evidence-based Practice Center.

Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California, June. Association for Computational Linguistics.

Patel, V., Branch, T., and Arocha, J. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *Int Journ Med Inform*, 65(3):193–211.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

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

Ronzano, F., Abura'ed, A., Espinosa-Anke, L., and Saggion, H. (2016). TALN at SemEval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016, San Diego, California, June. Association for Computational Linguistics.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

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

Rudell, A. P. (1993). Frequency of word usage and perceived word difficulty: Ratings of kuvera and francis words. *Behavior Research Methods, Instruments, & Computers*, 25:455–463.

Ruel, J., Kassi, B., Moreau, A., and Mbida-Mballa, S. (2011). *Guide de rédaction pour une information accessible*. Pavillon du Parc, Gatineau.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *ACL Student Research Workshop*, pages 103–109.

Sheang, K. C. (2019). Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 83–89, Varna, Bulgaria, September. INCOMA Ltd.

UNAPEI. (2019). *L'information pour tous*. UNAPEI.

Vajjala, S. and Meurers, D. (2015). Readability-based sentence ranking for evaluating text simplification. Technical report, Iowa State University.

Vander Stichele, R. (1999). Promises for a measurement breakthrough. In John Wiley & Sons, editor, *Drug regimen compliance. Issues in clinical trials and patient management*, pages 71–83. JM Metry and UA Meyer.

Vander Stichele, R. (2004). *Impact of written drug information in patient package inserts. Acceptance and benefit/risk perception*. Phd thesis, Ghent University, Ghent, Belgium.

Wani, N., Mathias, S., Gajjam, J. A., and Bhattacharyya, P. (2018). The whole is greater than the sum of its parts: Towards the effectiveness of voting ensemble classifiers for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 200–205, New Orleans, Louisiana, June. Association for Computational Linguistics.

Williams, M., Parker, R., Baker, D., Parikh, N., Pitkin, K., Coates, W., and Nurss, J. (1995). Inadequate functional health literacy among patients at two public hospitals. *JAMA*, 274(21):1677–1682.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zaharia, G.-E., Cercel, D.-C., and Dascalu, M. (2020). Cross-lingual transfer learning for complex word identification. *32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE*, 10.

Zampieri, M., Tan, L., and Genabith, J. (2016). Macsaar at semeval-2016 task 11: Zipfian and character features for complex word identification. 01.

Zeng, Q. T., Kim, E., Crowell, J., and Tse, T. (2005). A text corpora-based estimation of the familiarity of health terminology. In *ISBMDA 2006*, pages 184–92.

Zheng, W., Milios, E., and Watters, C. (2002). Filtering for medical news items using a machine learning approach. In *Ann Symp Am Med Inform Assoc (AMIA)*, pages 949–53.