A comparative study of word embeddings and other features for lexical complexity detection in French

Aina Garí[1] Marianna Apidianaki[1, 2] Alexandre Allauzen[1]
(1) LIMSI, CNRS, Univ. Paris Sud, Université Paris-Saclay, 91403 Orsay
(2) Computer and Information Science Department, University of Pennsylvania
aina.gari@limsi.fr, marianna@limsi.fr, allauzen@limsi.fr

RÉSUMÉ

Etude comparative de plongements lexicaux et autres traits pour la détection de la complexité lexicale en français

Détecter la complexité lexicale est une étape importante pour la simplification automatique de textes, servant lors de l'identification des éléments lexicaux à substituer. Dans ce travail, nous explorons l'utilité des plongements lexicaux pour mesurer la complexité de mots en français, en les combinant avec d'autres traits reconnus comme étant utiles pour cette tâche. Nos résultats sur une tâche d'ordonnancement de synonymes selon leur complexité montrent que les plongements seuls donnent de meilleurs résultats que nombreux autres traits, bien que leur performance reste inférieure à celle de systèmes basés sur la fréquence pour cette langue.

ABSTRACT

Lexical complexity detection is an important step for automatic text simplification which serves to make informed lexical substitutions. In this study, we experiment with word embeddings for measuring the complexity of French words and combine them with other features that have been shown to be well-suited for complexity prediction. Our results on a synonym ranking task show that embeddings perform better than other features in isolation, but do not outperform frequency-based systems in this language.

MOTS-CLÉS : Complexité lexicale, lisibilité, ordonnancement de synonymes, plongements lexicaux.

KEYWORDS: Lexical complexity, readability, synonym ranking, word embeddings.

1 Introduction

Complex word identification (CWI) is an important step in text simplification systems which aim at modifying the complexity level of texts to make them more accessible to readers with reading difficulties. There are many factors that can make a text complex or difficult, at the syntactic level (e.g. texts using passive voice), but also at the pragmatic and lexical levels (e.g. texts with a high number of domain-specific or rare words). Lexical complexity is the property of words that may pose comprehension difficulties to some readers, and which would not be part of the vocabulary of young children or low-proficiency foreign language learners. Complex words have a high impact on the readability of a text. Arya *et al.* (2011) showed that lexical complexity affected comprehension of elementary science texts by children, and manual lexical simplification has been shown to improve understanding and reading speed in dyslexics (Rello *et al.*, 2013). CWI can serve to identify the lexical elements in a text that need to be substituted by simpler synonyms or paraphrases for the meaning of the text to become more accessible.

We carry out a comparative study of different types of systems for measuring the complexity of French words. We experiment with features that have been identified as good indicators of complexity in the literature and word embeddings which, although successful in numerous semantics-related tasks, have not yet been applied to lexical complexity prediction. We assume that complex words appear in complex contexts, and simple words occur in texts that are easier to understand. We consider lexical complexity to be a property of words that is reflected in their context of use. Word embeddings encode rich contextual information, so we expect them to be particularly useful for complexity detection.

We train our systems on two French lexical resources that encode the distribution of words across different levels of difficulty : Manulex, where categories correspond to school levels (Lété *et al.*, 2004), and the FLElex database where words are categorized into second language proficiency levels (François *et al.*, 2014). We evaluate our systems on lists of synonyms that have been manually ordered by complexity.

2 Related Work

Due to the importance of lexical complexity detection for simplification, there have been numerous approaches to automatically predicting complexity. The task is often regarded as a classification problem into two or more levels of difficulty. Shardlow (2013) performs binary classification of words into simple or complex by means of a Support Vector Machine (SVM) classifier based on features such as frequency, number of syllables and senses. Gala *et al.* (2014) classify French words into three or six complexity levels, which indicate the distribution of words in the three school levels and six second language proficiency levels found in Manulex and FLElex, respectively. Gala *et al.*'s (2014) systems achieve accuracies close to 63% and 43% when trained and evaluated on the two resources. Interestingly, both Shardlow's (2013) and Gala *et al.*'s (2014) systems obtain results that are close to those of baseline models based on frequency only, and show that complex words occur more rarely in texts.

Another approach proposed by François *et al.* (2016) consists in ordering synonyms by complexity, creating thus a ranking instead of a classification. Their system is trained on pairs of French synonyms based on a complexity score derived from Manulex, and uses 21 features. We use in this study three of their best performing features : number of characters, number of phonemes and frequency.

Other works have focused on creating readability lexicons. Kidwell *et al.* (2009) develop a model that infers the age of aquisition of a word from its presence in graded texts and use it to assess text readability. Brooke *et al.* (2012) build a readability lexicon based, among others, on word co-occurrence information, which is the closest feature to the word embeddings we use in our experiments.

In the SemEval 2012 shared task on lexical simplification (Specia *et al.*, 2012), where systems had to choose a simpler synonym for words in context, the frequency baseline was very powerful and was only beaten by one system using psycholinguistic features including concreteness, imageability, familiarity and age of acquisition (Jauhar & Specia, 2012). When investigating the characteristics of lexical complexity for Spanish, Drndarević & Saggion (2012) conclude that combining frequency

Manulex

Class	Original	Balanced
1	9,384	9,384
2	8,040	8,040
3	19,421	9,000
Total	36,845	**26,424**

FLElex

Class	Original	Balanced
1	3,617	2,000
2	2,470	2,000
3	3,661	2,000
4	1,178	1,178
5	1,561	1,561
6	455	455
Total	12,942	**9,194**

TABLE 1 – Number of words available in the original Manulex and FLElex databases and in the balanced datasets. Words are classified per first class of appearance in each resource.

and word length is the best approach for choosing a substitute for a complex word.

3 Data and Resources

We train and evaluate our systems on two large lexical resources for French, **Manulex** and **FLElex**.

Manulex (Lété *et al.*, 2004) is a lexical database with word frequency information calculated from 54 French textbooks used in three school levels, with a total of 1.9 million word tokens. The levels correspond to ages of 6 (CP), 7 (CE1) and 8 to 11 (CE2-CM2) years old. We use its non-lemmatized version, which contains 48,886 word forms, and keep only unigrams with open parts of speech (nouns, adjectives, verbs and adverbs), which reduces it to 39,839 words.

FLElex (François *et al.*, 2014) is a graded lexicon based on textbooks and simplified readers for learners of French as a foreign language which contains a total of 777,000 tokens. This resource is organized in 6 levels corresponding to the proficiency levels of the CEFR scale (Conseil de l'Europe, 2001), ranging from A1 to C2. The resource contains 14,236 lemmas from which we retain 13,250 after filtering for part of speech and multi-words, as for Manulex.

We perform feature extraction and remove words for which no word embedding is available, which leaves us with 36,845 and 12,942 words in each dataset. Since words might occur in different levels, we assign each word in its first level of appearance, indicating its moment of acquisition, and we observe that the distribution of words is biased towards the third level in Manulex and the three first levels in FLElex. We believe this is because the third level in Manulex groups more ages than the first two, and in the case of FLElex the bias probably reflects the need of constructing a substantial lexicon during the first stages of learning. In order to learn from a more balanced dataset, we randomly remove words from the majority classes, obtaining a total of 26,424 words for Manulex and 9,194 for FLElex. The contents of the original and balanced resources are given in Table 1. 95% of the remaining data of each database is used for training, and 5% is used during the development stage to assess the correlation of the system's predictions with respect to the original classes the words belong to. To better estimate the complexity prediction capacity of the models on unseen data, we also remove from the training data words present in the synonym lists used for evaluation.

For evaluation, we use sets of synonyms manually ranked for complexity and made available by François *et al.* (2016). One example is *osseux* → *squelettique* → *maigre*, where words are ordered from more complex to simpler. Forty annotators were asked to order the synonyms according to their reading and comprehension difficulty without any given context. The resource consists of 36 groups of synonyms and 134 lexical units, with a few words being present in more than one group, ranked according to the majority ranking amongst all annotators. The reported Krippendorff's α of 0.399 reflects the difficulty of this task even for humans. Because of the unavailability of word embeddings for multiword expressions and for some synonyms present in the dataset,[1] we remove 13 words and 3 groups of synonyms (for which there was none or only one synonym left after removing the words), which results in a dataset of 33 sets of synonyms and 121 words. 5 of these sets are used as a development set, which results in a test set of 28 synsets and 104 words.

We extract frequency and number of phonemes of lexical items to be used by our models from **Lexique3** (New *et al.*, 2001), a large resource with various kinds of lexical information (phonetic transcription, syllabic structure, number of morphemes, etc.) for 142,728 French words.

4 Model description

We view complexity as a continuous, rather than categorical, notion with highly complex/simple words situated at the two extremes of the spectrum. In order to induce a continuous representation from the three Manulex levels and the six FLElex levels, we implement a simple feed-forward neural network that has as learning objective a number corresponding to the first level of appearance of a word in the corresponding resource. This serves as an indication of the age level or moment of acquisition of the word. The network takes as input features that represent the target word and outputs a real number that indicates its complexity. This way, we obtain a continuous representation of complexity from categorical data, which allows for a more fine-grained ranking than the one found in the original resources.

We experiment with different feature combinations, number of layers and layer sizes on a development set, and choose the best configuration for each feature combination for evaluation (see Section 5). We use three different architectures : one with no hidden layer, similar to logistic regression (no HL) ; one with a hidden layer of size 100 (1 HL) ; and one with two hidden layers of sizes 150 and 100 (2 HL). In all settings, rectified linear unit (ReLU) is used as an activation function and iterations are limited to 100.

For our experiments, we use the following lexical features which have been shown to be useful for predicting complexity in past work (Gala *et al.*, 2014) :

— **Number of characters**. The hypothesis is that longer words tend to be more complex, and shorter words tend to be simpler.

— **Number of phonemes**. This information is obtained from *Lexique3*, when available. For words not present in *Lexique3*, following Gala *et al.* (2014), we extract the information using eSpeak, an open source speech synthesizer.[2] The hypothesis is the same as for word length measured in terms of characters. We would expect the number of phonemes to be a better indication of word length for French given the abundance of many-to-one grapheme-phoneme correspondences.

— Log **frequency** in a corpus of film subtitles, as encoded in Lexique3. Frequencies in Lexique3 were calculated by averaging the frequency of appearance of a word per million counts in

[1] The lexical items with no embedding were : *agent de police, droit de grâce, grâce presidentielle, mine de crayon, mine antichar, mine antipersonnel, descente en rappel, empourprer, cramoisir, vagir.*

[2] http ://espeak.sourceforge.net

Features	Model	Spearman's correlation (ρ)	Pair-based	Rank-based (exact)	Rank-based (±1)
All features	no HL	**0.616**	0.786	0.490	**0.929**
Characters	no HL	0.307	0.734	0.375	0.680
Frequency	no HL	0.493	**0.812**	**0.606**	0.908
Phonemes	no HL	0.304	0.597	0.231	0.480
Embeddings	no HL	0.516	0.747	0.462	0.857
Freq + emb	no HL	**0.604**	0.773	0.471	0.908
Freq + char	1 HL	0.513	0.805	0.577	**0.929**
Freq + phon	1 HL	0.520	**0.812**	**0.625**	0.908
François et al. (2016)	-	-	0.789	0.596	0.902
Baseline	-	-	**0.820**	0.595	**0.935**

TABLE 2 – Results of the best systems trained on Manulex for each feature combination. Pair- and rank-based accuracy is measured on the manually ranked set of synonyms used for evaluation. Correlation is calculated against the ranking of these words in Manulex. The two best scores for each measure are marked in bold.

French subtitles of 4 types of films and series differing in their original language. Word frequency has been shown to be a very effective predictor of complexity in past work.

— 300-dimensional **word embeddings**[3] previously trained on Wikipedia with fastText using the Skip-Gram model (Bojanowski *et al.*, 2016). Word embeddings encode distributional information of words, and we expect them to be able to capture differences in complexity. Our hypothesis is that complex words appear in more complex texts.

5 Evaluation

We evaluate our systems on the manually ordered synonym lists described in Section 3 measuring the **pair-based** and **rank-based** accuracy of the predictions against human judgments. For the pair-based evaluation, we exhaustively create complex → simple pairs of synonyms from every set of synonyms. For example, from the group *forger → formuler → former → inventer*, ordered from complex to simple, we derive 6 pairs : (*forger → formuler, forger → former, forger → inventer, formuler → former, formuler → inventer, former → inventer*), preserving the directionality indicated by the annotators. We expect a good system to assign higher scores to more complex words, producing the correct ordering. We report the proportion of pairs that were correctly ordered by the system.

In the **rank-based** evaluation, the system has to order a full group of synonyms based on the score that it assigns to each word. We report the percentage of cases in which the system correctly predicts the exact position of a synonym. Following François *et al.* (2016), we additionally report the proportion of cases where a synonym was placed in the original place or only one position away. For calculating the latter, we exclude groups of synonyms with only two words, which would otherwise always count as correct (there is only one such pair in the development set and one in the test set). In the case of ties between the scores of two synonyms, they are considered to be incorrectly placed, since this means the system is not able to detect the difference in complexity between the two words.

We test the three proposed architectures (with 1 or 2 HL(s), and without HL) on the development sets extracted from Manulex and FLElex and from the manually-ranked synonym resource. We retain as

[3] Available for 294 languages at https ://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Features	Model	Spearman's correlation (ρ)	Pair-based	Rank-based (exact)	Rank-based (± 1)
All features	1 HL	0.429	0.701	0.519	0.816
Characters	1 HL	0.291	0.734	0.375	0.694
Frequency	no HL	**0.569**	0.812	**0.606**	0.908
Phonemes	no HL	0.275	0.597	0.231	0.500
Embeddings	no HL	0.499	0.753	0.558	0.837
Freq + emb	2 HL	0.541	0.688	0.471	0.786
Freq + char	1 HL	**0.596**	0.805	**0.606**	**0.939**
Freq + phon	1 HL	**0.596**	**0.825**	**0.615**	0.908
François et al. (2016)	-	-	0.789	0.596	0.902
Baseline	-	-	**0.820**	0.595	**0.935**

TABLE 3 – Results of the best systems trained on FLElex for each feature combination. Pair- and rank-based accuracy is measured on the manually ranked set of synonyms used for evaluation. Correlation is calculated against the ranking of these words in FLElex. The two best scores for each measure are marked in bold.

Freq + char (correct)	dépouiller → dérober → piquer → voler
Embeddings	dérober → dépouiller → piquer → voler
Freq + char	mental → spirituel → fin
Embeddings (correct)	spirituel → mental → fin

TABLE 4 – Examples of synonyms that were correctly and incorrectly ranked by one of the best performing systems (Freq + char) and the embeddings-based system, both trained on FLElex.

the best model for each feature combination the one that obtains the strongest correlation (Spearman's ρ) with Manulex and FLElex original classes and the highest accuracy scores in the pair-based and rank-based evaluations on the synonym development set. In the case of two models performing equally or very similarly, we choose the simplest one (i.e. the one with fewer layers).

We compare the results of the best model for different feature combinations to the ones produced by François *et al.*'s model on the same test set. The size of the dataset used in our evaluation is slightly different from the one used in their original work, since lexical items that had no embeddings were left out. We recompute the accuracy of their system output on the reduced test set. As a baseline, we use the simple frequency feature with no training involved, which considers more frequent words to be simpler and orders them accordingly.

6 Results and discussion

Results are presented in Tables 2 (Manulex) and 3 (FLElex). The patterns observed and the scores obtained by the systems trained on Manulex and FLElex are, in general, similar. One could expect FLElex to result in a system producing better rankings because of its finer granularity with respect to Manulex, but the much bigger size of the latter probably counteracts this effect.

Among all feature combinations, frequency combined with number of characters or with phonemes seem to be the best predictors of complexity for both Manulex and FLElex. The number of phonemes is the worst performing feature when used on its own, and obtains the lowest score in all measures. This is somewhat surprising, since Gala *et al.* (2014) show that phonemes are slightly more strongly correlated to Manulex and FLElex classes than characters, which do a bit better in this setting but still are the next poorest performing feature. However, these features alone produce a considerable amount of ties which substantially lower their performance.

The combination of word embeddings with a strong feature such as frequency does not seem to improve the results of embeddings alone on FLElex. The system that relies only on embeddings obtains better results on this dataset than the system using all features. On Manulex, on the contrary, embeddings seem to benefit from being combined with other features. In Table 4 we give an example of a group of synonyms that is correctly ranked by one of the best systems – the one that combines frequency and length in characters (Freq+char) – in the FLElex evaluation, and which the embeddings-based system ranked incorrectly. The lower part of the table shows a case where the opposite happened. We observe that the systems switch words that are situated next to each other in the rankings and which would be more difficult to rank than words situated in the two extremes, given that they might be similarly complex or simple. The difficulty of the task is highlighted by the low inter-annotator agreement reported by François *et al.* (2016) which must be partly due to this type of hard-to-rank cases.

The frequency-based baseline is very powerful and is only beaten with a small margin in the rank-based accuracy evaluation by more complex models that have frequency among their features. François *et al.* (2016)'s system is one of the best performing models, but whilst it employs 21 different features, its accuracy is slightly below that of simpler frequency-based models. This adds more evidence to the established idea that frequency is a crucial indicator of lexical complexity.

7 Conclusions

In this work, we have explored the use of word embeddings alone and combined with other features for lexical complexity prediction in French. Our system learns a complexity score from complexity levels based on age of acquisition (Manulex) and second language proficiency (FLElex). The evaluation on a synonym ordering task seems to indicate that whereas word embeddings obtain better results than other features in isolation, frequency-based systems – even a simple frequency baseline – are better suited for this task. The best performing systems use frequency or combine it with characters or phonemes, two features that have been proven to be useful in past work. In the future, we plan to experiment and compare with other kinds of word embeddings built from corpora with a higher level of stylistic variation.

8 Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful and constructive comments. We particularly would like to thank Núria Gala and Thomas François for the interesting discussions on French lexical complexity resources, and for sharing with us the manually annotated synonyms dataset used for evaluation. This work has been supported by the French National Research Agency under project ANR-16-CE33-0013.

Références

ARYA D. J., HIEBERT E. H. & PEARSON P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Elementary Education*, **4**(1), 107.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

BROOKE J., TSANG V., JACOB D., SHEIN F. & HIRST G. (2012). Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, p. 33–39.

CONSEIL DE L'EUROPE (2001). *Cadre européen commun de référence pour les langues*. Paris : Didier.

DRNDAREVIĆ B. & SAGGION H. (2012). Towards automatic lexical simplification in spanish : an empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, p. 8–16.

FRANÇOIS T., BILLAMI M., GALA N. & BERNHARD D. (2016). Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL 2016*, volume 2, p. 15–28.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded lexical resource for french foreign learners. In *LREC*, p. 3766–3773 : Citeseer.

GALA N., FRANÇOIS T., BERNHARD D. & FAIRON C. (2014). A model to predict lexical complexity and to grade words (un modèle pour prédire la complexité lexicale et graduer les mots)[in french]. *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, **1**, 91–102.

JAUHAR S. K. & SPECIA L. (2012). Uow-shef : Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 477–481.

KIDWELL P., LEBANON G. & COLLINS-THOMPSON K. (2009). Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, p. 900–909.

LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). MANULEX : A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*, **36**(1), 156–166.

NEW B., PALLIER C., FERRAND L. & MATOS R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE^{TM}//a lexical database for contemporary french : LEXIQUE^{TM}. *L'année psychologique*, **101**(3), 447–462.

RELLO L., BAEZA-YATES R., DEMPERE-MARCO L. & SAGGION H. (2013). Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, p. 203–219 : Springer.

SHARDLOW M. (2013). A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, p. 103–109, Sofia, Bulgaria.

SPECIA L., JAUHAR S. K. & MIHALCEA R. (2012). Semeval-2012 task 1 : English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 347–355.