# CEFR-based Contextual Lexical Complexity Classifier in English and French

**Desislava Aleksandrova** and **Vincent Pouliot**
CBC/Radio-Canada
dessy.aleksandrova@radio-canada.ca, vincent.pouliot@radio-canada.ca

## Abstract

This paper describes a CEFR-based classifier of single-word and multi-word lexical complexity in context from a second language learner perspective in English and in French, developed as an analytical tool for the pedagogical team of the language learning application *Mauril*. We provide an overview of the required corpora and the way we transformed it into rich contextual representations that allow the disambiguation and accurate labelling in context of polysemous occurrences of a given lexical item. We report evaluation results for all models, including two multi-lingual lexical classifiers evaluated on novel French datasets created for this experiment. Finally, we share the perspective of Mauril's pedagogical team on the limitations of such systems.

## 1 Introduction

The lexical complexity classification task exists in its simplest form as the binary complex word identification task (CWI) and at its most complex, as a multi-class classification where the class nomenclature and granularity is determined by the labelling of the training data. Lexical complexity finds application in text and lexical simplification systems, in automated language proficiency assessment, as well as in the creation of level-appropriate pedagogical content, which also happens to be the use case of *Mauril* [1].

*Mauril* is a new, free digital platform leveraging a wide range of stimulating and entertaining content from CBC and Radio-Canada to help users learn English and French. Financed and endorsed by the Government of Canada, this new tool is designed and deployed by CBC/Radio-Canada, in collaboration with a committee of pedagogical experts. It's meant to help improve oral comprehension and integrate language knowledge in everyday life.

The language learning process in Mauril begins with a placement test and is then organized by levels covering beginner, intermediate and advanced proficiency [2] . Each level contains units and each unit consists of a video clip of varying length (anywhere from 1 to 22 min) punctuated by comprehension questions and accompanied by highlighted vocabulary (words and expressions) with a corresponding difficulty level.

The creation of pedagogical content (from the selection of video segments, through questions and vocabulary definition to the assignment of difficulty level) is performed manually by experienced foreign language teachers. This labour intensive process of content creation was the target of a lexical processing pipeline designed to streamline and facilitate the extraction and addition of more level-appropriate vocabulary to all existing units. The system in question had to be able to parse the subtitle file of a video segment, reconstruct and then segment the text into tokens, detect and extract multi-word expressions and then assign a complexity label to all occurrences of words and expressions in context. The central component of this system and the current publication is a CEFR-based [3] lexical complexity classifier for both French and English.

In this paper, we apply a novel approach to lexical complexity prediction (LCP), based on rich contextual representations. We show that our system is capable of:

- classifying word senses in context;
- predicting complexity of both words and phrases;
- producing results in French with no or limited training data

[1] https://mauril.ca/en/

[2] cf. § 5.1 for a mapping between Mauril's levels and other standards for language ability assessment.

[3] The *Common European Framework of Reference (CEFR)* is a common basis for the elaboration of pedagogical materials and an international standard for describing the proficiency of foreign-language learners (Council of Europe, 2001).

*Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527
July 13, 2023

## 2 Related work

In the context of their language-learning platform offering a digital language proficiency assessment exam, DuoLingo had developed and released a CEFR checker (now discontinued) allowing users to validate the difficulty of words and text in English and Spanish. The lexical complexity component of the tool was described in Settles et al. (2020) as a CEFR-based vocabulary scale model based on CEFR vocabulary wordlist (an inventory of 6,823 English words labelled by CEFR level, mostly in the B1/B2 range). The authors proposed two regression models fit on lexical item representations composed of surface-based features aimed as a proxy of frequency. The models did not seem to handle multi-word expressions, nor common contractions such as *doesn't* and *you've*. Their complexity predictions were lemma-based and did not take inflection into account, which was evident and consequential in Spanish more than it was in English. Finally, the misclassifications reportedly attributed to polysemy (Settles et al., 2020, p.6) were in fact cases of homonymy, since the representations did not include PoS information.

Disambiguating polysems is a challenge for all lemma-based complexity corpora (FLE, 2004; Lété, 2004; Cobb, 2007; Lonsdale and Le Bras, 2009; François et al., 2014; Schmitt et al., 2021) which conflate polysemous entries into a single entry and assign it a single level. However, not all senses of a polysemous word are learned at once and the different meanings of polysemous words are not uniformly distributed across texts of varying difficulty. Francois and Watrin (2011) even found a negative association between frequency and complexity with more frequent words being associated with more complex texts. This may be attributable to the fact that frequent words tend to be more polysemous (Zipf, 1945) and complex texts are likely using more than one of those meanings disguised as occurrences of the same lemma. In fact, learners encounter highly polysemous words most often (Crossley et al., 2010), hence the importance of disambiguating and accurately predicting the complexity of word senses.

The role of context in LCP is two-fold. Firstly, it is crucial in deriving the correct sense of a polysemous word (word sense disambiguation), as words in isolation provide no information as to their intended meaning. Secondly, it has an incidence on a word's complexity as a source of complementary information. Learners acquire much of their vocabulary knowledge from context rather than from decontextualized forms such as word lists, definitions, etc.) (Nagy, 1995) Gooding and Kochmar (2019a) were some of the first to recognize the importance of context for the task of CWI. As they correctly pointed out, the perceived complexity of the lexeme *molars* in the phrase *Elephants have four molars...* may be higher than in the phrase ... *new molars emerge in the back of the mouth.* since the second occurrence is surrounded by familiar words that imply its meaning, while the first co-occurs with the rarer and less semantically similar *elephants*.

In more recent work, (Alfter and Volodina, 2018; Alfter, 2021) found that one of the most important predictors of complexity in their experiments was topic distribution – a context feature modelling polysemy and defined as a vector indicating all topics under which a word occurred. Effects of the inclusion of context on predicting lexical complexity are also discussed in a recent survey of LCP by (North et al., 2023).

In contrast, lexical complexity work on French has mostly focused on representing and classifying lexical items in isolation (Gala et al., 2013; François et al., 2016), independently of the context in which they appear. This position is reflected in the lexical complexity corpora available in French (François et al., 2014; Lété, 2004) which provides no means of contextualizing or disambiguating word senses. Gala et al. (2014) have presented lexical classification models trained on these corpora where lexical items were represented by 49 orthographic, morphologic and statistical features. Their L2 classifier achieved 43% accuracy on the six-way classification task.

Approaches based on such linguistic features often struggle to represent MWEs since the latter are absent from vocabulary lists despite their high frequency in everyday interactions [4] and invite the use of simplifying techniques such as averaging the constituents of the MWE (which in turn wrongly assume compositionality). At the same time, studies in both French and English have shown the importance of MWE-based features for the accurate assessment of lexical complexity (Francois and Watrin, 2011; Kochmar et al., 2020).

[4] Jackendoff (1995) estimates that not less than half of the lexical units readily available to a speaker in daily interactions are MWEs.

## 3 Training data

To train a contextual classifier of lexical items, we needed a collection of words and expressions associated with complexity levels and accompanied by at least one sentence illustrating their usage in context.

### 3.1 English

For the English classifier, we used the Cambridge University Press's English Vocabulary Profile [5] (Capel, 2010, 2012), following Settles et al. (2020). The EVP corpus is a rich resource in British and American English which associates single words, phrasal verbs, phrases, and idioms (Table 1) not only with a CEFR level and a part of speech tag (PoS), but with a definition, a dictionary example and production examples on the basis of several hundred thousand examination scripts written by learners from all over the world. It offers reliable information about which words (and more importantly, which meanings of those words) ARE known and used by learners (rather than SHOULD be known) at each level of the CEFR. For example, we find 10 entries for the word form *run* in the American English section of the corpus, two noun forms and eight verbs, whose complexity varies between A1 (*He can* ***run*** *very fast.*) and C2 (*He would like to* ***run*** *for mayor.*) Each of those meanings is accompanied by usage examples taken from essays of students whose acquisition level corresponds to the complexity level of the word. Such contextual examples allowed us to include disambiguated polysemous lexemes with varying complexity to the training data.

| Word form | POS | Level |
|---|---|---|
| sleep | verb | A1 |
| sleep with sb | phrasal verb | C2 |
| lose sleep over sth | idiom | C2 |
| not sleep a wink | phrase | C2 |

Table 1: Example entries from the EVP corpus

After extracting all triplets `<word form, level, examples>` from the American subset of the corpus, we made sure that each word form's inflection matches the inflection of its occurrence in at least one usage example. Those who differed were modified manually to assure such correspondence. Uninflected phrases such as *not sleep a wink* became *didn't sleep a wink* to include the auxiliary verb present in the usage example. Phrasal verbs and expressions with placeholder arguments such as *sleep with* ***sb***, *rush into* ***sth*** lost the arguments. Placeholder arguments in non-contiguous expressions such as *grab* ***sb's*** *attention* were replaced by actual arguments from the entry's usage examples: *grab the reader's attention*, *grab people's attention*. Complex items with word order variation such as *set back sb/sth or set sb/sth back* were split into multiple word forms. Following the edits, the dataset contained 14,177 entries distributed unevenly in six classes (Table 2) of which 90% were used for training and the remaining 10% were kept for evaluation.

| A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|
| 804 | 1525 | 2715 | 3829 | 2159 | 3145 |

Table 2: Class distribution of the EVP corpus

### 3.2 French

To our knowledge, no lexical complexity corpora in French resembles *EVP* and its disambiguated, contextualized, CEFR labelled words and expressions extracted from ESL production corpora.

*FLELex* (François et al., 2014) is a graded lexicon for French as a foreign language (FFL) that reports the normalized frequencies of words (lemmas) across CEFR levels. The frequency distributions have been estimated on a corpus of FFL textbooks and FFL simplified books rather than on learners' corpora. Polysemous lemmas in FLElex are ambiguous and conflate the frequencies of all lexemes with the same spelling. As a consequence, the associated frequency distribution is likely right-skewed, reflecting the relatively higher frequencies of easier meanings. In addition to lacking usage examples, it requires a mapping between frequency distributions and CEFR classes (Gala et al., 2013; Alfter et al., 2016; Pintard and François, 2020).

*Manulex* (Lété, 2004) offers frequency distributions of 23K+ French lemmas and 43K+ word forms across three primary school levels rather than CEFR. As a further limitation, the corpus contains no usage examples.

*A Frequency Dictionary of French: Core Vocabulary for Learners* (Lonsdale and Le Bras, 2009) enumerates the 5000 most frequent lemmas with a usage example in French and absolute frequency among other attributes. Word frequency, how-

[5] https://www.englishprofile.org/wordlists/evp

ever, is a necessary but not sufficient predictor of lexical complexity. *LexTutor*'s frequency lists (Cobb, 2007) contain only lemmas and no complexity labels or usage examples. *Les référentiels* (FLE, 2004) compile lexical inventories across most CEFR levels based on target competence rather than on actual learner performance. Very few lemmas are paired with a sentence, but an organization by themes makes it possible to manually disambiguate homographs and polysemous words within and across complexity levels.

## 4 Evaluation data

### 4.1 English

To evaluate the classifiers we trained on English data, we used 10% of the EVP corpus. This approach to evaluation, despite being methodologically sound, has a tendency to overestimate performance since evaluation and training data have the same distribution.

### 4.2 French

To evaluate the models' performance on French, we had to create labelled data since none was readily available. The current section describes three versions of our French Evaluation Corpus (FEC) two of which are based on parts of *Les référentiels* (FLE, 2004), a series of word indexes that serves as a lexical reference for learners from levels A1 to C1. Each level is subdivided into chapters that, in turn, break the lists of words and expressions into different themes. Some lexical items are accompanied by context sentences.

We first transcribed 13,016 words and expressions (for levels A1 to B2) with their corresponding PoS and examples, whenever available. Since the advanced level C1 was out of scope for Mauril's use case, we only extracted 10 examples from it. We then excluded all vocabulary belonging to European varieties of French (e.g. *atriaux, boutefas, longeole, schublig*, all types of sausages from Switzerland) and only kept lexical units actively used in Québec. We also identified and erased many duplicate entries through and across levels, while manually disambiguating and preserving occurrences of polysems. The last systematic edit we made to the list was to omit MWEs that were considered non-productive or redundant.

For the first version of the corpus (FEC1), we kept only entries which already had context examples for a total of 914 (Table 3). Despite the extensive cleaning and editing, we found that the corpus still contained many inconsistencies which motivated the creation of other versions.

| | | |
|---|---|---|
| contre | B1 | Mets cette chaise **contre** le mur. |
| *against* | | *Put this chair **against** the wall.* |
| contre | B1 | Je suis **contre** son projet. |
| *against* | | *I am **against** his project.* |

Table 3: Examples from the FEC1 corpus

For the second version (FEC2), we extracted the lexical items from several themes (semantic fields) across all levels, making sure to avoid the contradictions present in the first version by not allowing multiple occurrences of the same lexeme (at any level). The resulting list contained 473 lexical items (A1: 83, A2: 99, B1: 114, B2: 167, C: 10) most of which did not have a corresponding example in *The Référentiels*. We had usage examples created for all lexemes by a trained linguist, native speaker of French. Since sentences were aimed to be understandable in isolation, most of them followed a simple, declarative SVO structure with very few having subordinate clauses, complex noun phrases or non-pronominal subjects. Still, we were unable to make sure that the sentence complexity of each example is equal or lower to the complexity level of the lexical item whose usage it aimed to illustrate (the way a performance corpus such as *EVP* naturally does).

The third version of the corpus (FEC3) is based on a series of FSL [6] textbooks from Quebec (Gouvernement du Québec, 2014) covering levels A1 to B2 and targeting adult learners (Table 4). By extracting vocabulary (in context) from listening and reading comprehension activities, we could better control for the difficulty of the usage examples, even though the resulting complexity labels still equate comprehension rather than production. FEC3 is the smallest and most *Quebecois* corpus of the three with 48 lexical items in each of the 4 levels. While compiling the corpus, we noticed that it contained more advanced words taught at lower levels than the previous source. We attribute it to the didactic materials being developed following the FLI [7] approach and targeting adults integrating a new country.

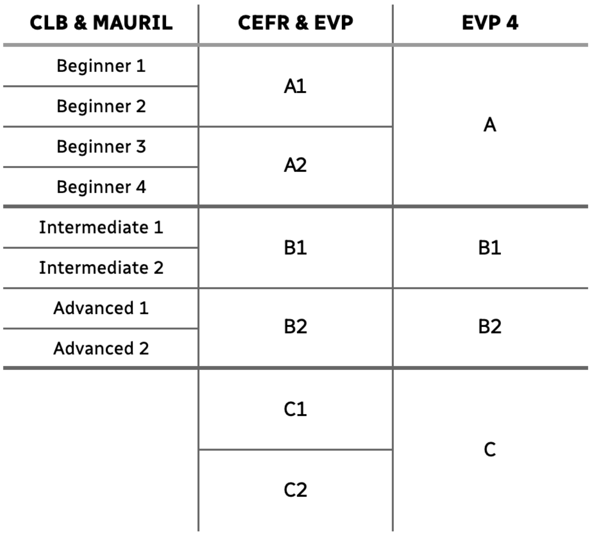
All three versions of the FEC reflect competence rather than performance, contrary to the training

[6] French as a second language

[7] French Language of Integration

| | | |
|---|---|---|
| imbibez | A2 | **Imbibez** un linge de vinaigre chaud ou froid. |
| *soak* | | ***Soak** a cloth in hot or cold vinegar.* |
| compte | A2 | Si vous disposez de fonds dans votre **compte**, vous pouvez envoyer de l'argent dans le monde entier. |
| *account* | | If you have funds in your **account**, you can send money worldwide. |

Table 4: Examples from the FEC3 corpus



| CLB & MAURIL | CEFR & EVP | EVP 4 |
|---|---|---|
| Beginner 1 | A1 | A |
| Beginner 2 | | |
| Beginner 3 | A2 | |
| Beginner 4 | | |
| Intermediate 1 | B1 | B1 |
| Intermediate 2 | | |
| Advanced 1 | B2 | B2 |
| Advanced 2 | | |
| | C1 | C |
| | C2 | |

Figure 1: Class mappings between CLB & Mauril levels, the six levels of CEFR & EVP, and the rebinned version of the EVP corpus with four classes

data used to create the classifier.

## 5 Lexical classification

In this section, we describe the creation of a classifier able to assign a complexity level between $1 \equiv A1$ and $6 \equiv C2$ to the meaning of any word or multi-word expression as determined by its context.

### 5.1 Classes

Lexical complexity may be cast as a 6-class classification problem whenever training data is available for all CEFR levels. Mauril's pedagogical content is distributed among eight levels and covers two of the three proficiency stages defined in the Canadian Language Benchmark's nomenclature: Basic and Intermediate [8] . These eight levels correspond to four of the CEFR levels, as illustrated in Figure 1. Given the relatively small number of examples in A1 and A2 (cf. Table 2) and Mauril's coverage, the lexical classification need of Mauril is better satisfied by a 4-class classifier with a combined class for both the beginner and the advanced levels. In this way, each of the beginner, intermediate and advanced levels in Mauril corresponds to a class label with the fourth label C covering advanced vocabulary beyond the current pedagogical scope of the application (Figure 1).

### 5.2 Preprocessing

The minimal preprocessing of the triplets targets the word form and the examples and consists of tokenization using spaCy's models for English and French [9]. An additional preprocessing step of expanding some common unambiguous contractions in English (e.g. *don't* → *do not*) improves the tokenization.

### 5.3 Vectorization

Rather than representing the vocabulary items by their frequency and/or surface-level characteristics (e.g. number of characters, number of syllables, etc.), we obtain a semantic, contextual, dense vector representation of each item from a pre-trained masked language model (Devlin et al., 2018).

Unlike word2vec models which are sources of non-contextualized (or static) embeddings, trained masked language models such as BERT assign a different representation to each instance of a word in a different context. Garí Soler and Marianna Apidianaki (2021) showed that nonetheless, such language models encode information about a word's monosemous or polysemous nature. Their experiments also showed that the uncased BERT model possessed more knowledge about lexical polysemy than the cased one.

To obtain a vector representation reflecting a particular meaning of a string, we encode (using the model bert-base-uncased [10]) each of the usage examples of a triplet containing the string and then select the WordPieces[11] composing it. For each WordPiece, we extract and sum the vector representations from the 12 hidden layers. Finally, we aggregate the vectors of all WordPieces by averaging them. When more than one usage examples

[8] The two Advanced levels in Mauril correspond to CLB's levels 7 and 8, both belonging to the Intermediate proficiency stage

[9] https://spacy.io/ | v. 3.1.3 | en-core-web-lg, fr-core-news-lg

[10] https://huggingface.co/bert-base-uncased

[11] Subwords resulting from a segmentation algorithm

are accompanying a word form, we take the mean of all occurrences as a final representation. We considered different selection and pooling strategies for the hidden layer representations: first, last, second-to-last layers, summing or concatenating the last four hidden layers. The SVC model trained on the sum of all 12 hidden layers achieved the highest accuracy in a 3-fold cross validation.

In this manner, embeddings of tokens with the same or with similar meanings are more alike (in terms of cosine similarity) despite the varying context, than embeddings of homographs with unrelated senses.

Table 5 illustrates eight occurrences of the token *run* in contextual minimal pairs, where each context evokes a different meaning (present in the EVP dataset). We compared the embeddings of the same token in each of the new contexts in Table 6 to the vectors in Table 5 to find the closest meaning in terms of pairwise cosine similarity. The experiment shows that as long as their contexts evoke the same meaning, the embeddings of two occurrences of the same word would remain very similar.

| # | WORD FORM IN CONTEXT | MEANING |
|---|---|---|
| 1 | I **ran** a marathon | MOVE FAST |
| 2 | I **ran** the program | OPERATE |
| 3 | I **ran** into trouble | ENCOUNTER |
| 4 | I **ran** into the kitchen | ENTER |
| 5 | I **ran** into a friend | MEET |
| 6 | I **ran** an ad | PUBLISH |
| 7 | I **ran** the water for 20 min | LIQUID |
| 8 | I **ran** for president | ELECTION |

Table 5: Minimal pairs of sentences illustrating different meanings of the word form *ran*

### 5.4 Classification

We used a support vector classifier algorithm [12] (Platt et al., 1999) with adjusted class weights inversely proportional to class frequencies in the input data to correct for the class imbalances. For the same reason, we calculate and report a balanced accuracy score [13] defined as the macro-average of recall scores per class.

### 5.5 Transfer learning

In the absence of appropriate training data in French, we used a transfer learning approach to train a multilingual lexical classifier. We replaced the monolingual source of embeddings with `bert-base-multilingual-uncased` [14] and trained a classifier on the new representations of the English training data. The resulting model is capable of encoding and classifying multilingual input, including in French by leveraging correlations present in the monolingual training data. We hypothesize that even though the lexicalization of senses and their associated complexity varies across languages, there are reliable regularities between form, meaning, and difficulty present in many languages, especially closely related ones (such as English and French). The accuracy of feature-based lexical classifiers has showed that between 40 and 65% of the variance in lexical complexity models can be explained by universal properties such as frequency, word length and other stylometric characteristics (Gala et al., 2014; Alfter and Volodina, 2018; Alarcon et al., 2019). Ideally, the approach of transfer learning should be applied from morphologically richer languages (such as French) to languages with less inflectional variability (such as English), provided training data is available.

| WORD FORM IN CONTEXT | CLOSEST MEANING | COS. SIM. |
|---|---|---|
| He **ran** 24 miles | MOVE FAST | 0.89 |
| She **ran** the race | MOVE FAST | 0.87 |
| You **ran** the script | OPERATE | 0.90 |
| He **ran** into problems | ENCOUNTER | 0.94 |
| The car **ran** into a pothole | ENCOUNTER | 0.88 |
| She **ran** for mayor | ELECTION | 0.92 |
| He **ran** for office | ELECTION | 0.92 |
| I **ran** into the house | ENTER | 0.99 |
| He **ran** into the president | MEET | 0.94 |
| I **ran** into you | MEET | 0.95 |

Table 6: The word form *ran* in different contexts with its corresponding closest meaning in terms of cosine similarity

## 6 Results and Discussion

To establish the effectiveness of feature-based representation as lexical complexity predictors on the EVP dataset, we trained the model `ME6 Baseline`, a support vector classifier (with `class_weight="balanced"`) fitted on frequency and two common surface features: the length of

[12] sklearn.svm.SVC

[13] sklearn.metrics.balanced_accuracy_score

[14] https://huggingface.co/bert-base-multilingual-uncased

the word form in characters and in tokens. Without contextual information, we could only disambiguate some of the homographs by part-of speech and had to reduce multiple occurrences of a single `token+POS` pair to the one with the lowest complexity level. This resulted in 11,133 data points of which we used 90% for training and 10% for evaluation. The resulting confusion matrix with normalized scores per class on Figure 2 shows poor recall for all inner classes, especially the A2 level.

We then trained a model called `ME6 Contextual` which fits the same support vector classifier on the 6-class training set of the EVP corpus (cf. §3.1) with word embeddings extracted from a monolingual language model (as described in §5.3). For this experiment, we could disambiguate and use all training points, including polysems, resulting in a larger training set (12,760). The evaluation on 10% of the corpus (1,418 data points) produced the results on Figure 3. The improved performance of the contextual model is consistent across all six classes and visible on both the confusion matrix as precision and recall and Table 7 in terms of F1 scores. Classification errors are limited to the neighbouring classes.

We further trained a 4-class classifier (`ME4 Contextual`) on a rebinned and rebalanced [15] version of the dataset. The reduced number of classes provides a further improvement of F1 scores (Table 7) despite the reduction in training data caused by the rebalancing.

To train the multilingual model `MME4 Contextual`, capable of classifying not only English but also French words and expressions, we used the method described in section 5.5. When evaluated on English data, the model performs almost as well as the one trained on contextual vectors from a monolingual BERT (Table 7). When evaluated on French data however (cf. § 4.2), there seems to be a significant drop in performance, most noticeable in the intermediate classes. The model overestimates the complexity of all classes by predicting the label C for the majority of examples which explains the poor F1 scores of the C class.

The last model we evaluated, `MMEFR4 Contextual`, was trained on a combination of English and French labelled data from EVP and FEC1. Despite the shortcomings of the evaluation corpora we produced in French, it could be used for training, especially the FEC1 which has 904 examples labelled from A1 to B2 [16] . After rebalancing the resulting joined dataset, we trained the same support vector classifier on 99% of the data, leaving 1% for evaluation.

The results listed in Table 7 show a significant improvement of the F1 scores for all classes (except for C where the complexity of the 10 examples in the evaluation sets is now underestimated) on FEC2 and FEC3. The scores on English data confirm that the gain in French was not achieved at the expense of the performance in English.

We will be releasing [17] code and English data used for training as well as trained models with the exception of any model trained on a combination of English and French data.

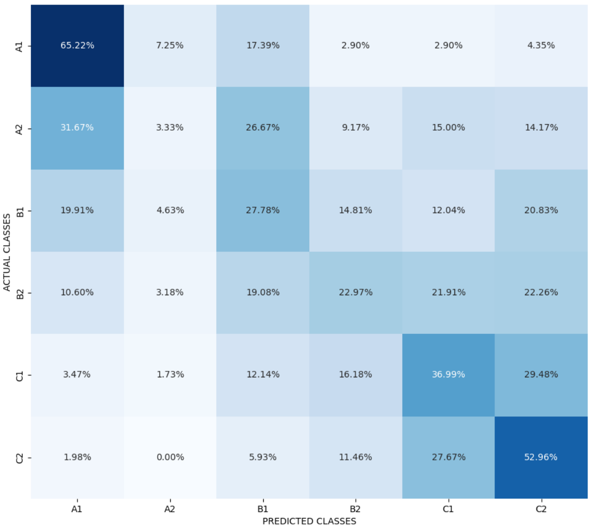


Figure 2: Confusion matrix with normalized scores per class of model ME6 Baseline

Analysis of the errors on the French evaluation corpora showed the significance of the context's complexity for the individual lexical item's complexity prediction. Naturally, contextual representations of lexical items are influenced by the surrounding words and syntactic structures, but the extend to which this affects the lexical classifier becomes more visible in a competence type of corpora such as FEC1-3. Furthermore, analysis of the errors on French corpora show that when the model has only seen English data, it has a tendency to overestimate the complexity of inflected French verbs since the training data does not reflect the

[15] To balance the classes, we reduced the size of the largest class C to the size of the second largest – B2.

[16] We excluded the ten examples of the C class since those are present in FEC2 and FEC3

[17] https://github.com/cbcrc/vocabclf

| Model | Lang. | A1 | A2 | B1 | B2 | C1 | C2 | Test Set |
|---|---|---|---|---|---|---|---|---|
| ME6 Baseline | en | 0.38 | 0.05 | 0.29 | 0.29 | 0.31 | 0.47 | 10% of EVP |
| ME6 Contextual | en | 0.63 | 0.49 | 0.45 | 0.50 | 0.42 | 0.60 | 10% of EVP |
| ME4 Contextual | en | 0.69 | | 0.42 | 0.53 | 0.71 | | 10% of EVP |
| MME4 Contextual | en, fr | 0.66 | | 0.39 | 0.51 | 0.70 | | 10% of EVP |
| | en, fr | 0.66 | | 0.13 | 0.17 | 0.05 | | FEC1 |
| | en, fr | 0.60 | | 0.15 | 0.06 | 0.06 | | FEC2 |
| | en, fr | 0.47 | | 0.18 | 0.13 | 0.12 | | FEC3 |
| MMEFR4 Contextual | en, fr | 0.65 | | 0.40 | 0.35 | 0.71 | | 1% of (EVP + FEC1) |
| | en, fr | 0.68 | | 0.31 | 0.51 | 0.00 | | FEC2 |
| | en, fr | 0.62 | | 0.27 | 0.30 | 0.00 | | FEC3 |

Table 7: Language support, F1 scores, and test set of trained models

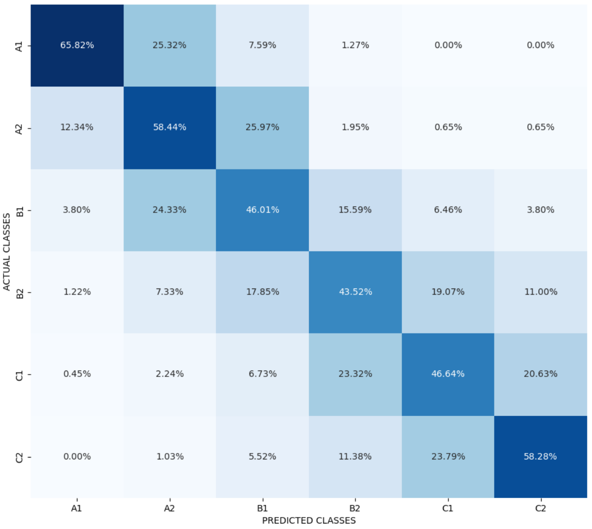


Figure 3: Confusion matrix with normalized scores per class of model ME6 Contextual

complex morphology of the French language. Still, the rich contextual representations encode enough information to allow the model to correctly distinguish and classify, for example, instances of the verb *faire* used as a main verb vs. as an auxiliary.

The early adoption and tests of the vocabulary processing pipeline by Mauril's team of foreign language teachers highlighted some of the models' limitations. For example, none of the models (not even the multilingual one) give a special treatment to cognates (sets of words in one of the two languages that have been inherited in direct descent from the other one) which are normally considered easier to acquire. Another concern has been the lack of transparency in the classifier's predictions, a direct consequence of the dense representations we favoured over the more interpretable linguistic features. Finally, a contextual classifier may predict different levels for occurrences of the same lexeme in different contexts. Those limitations underline the need for human validation of the output of such systems.

## 7 Conclusion

In this article, we detailed the creation and evaluation of a lexical complexity classifier in French and English, predicting contextually-aware CEFR-based labels for words and multi-word expressions alike. We established a baseline for the six-way lexical classification on the EVP corpus and showed that replacing the representation by statistical features such as frequency for a dense contextual embedding from a masked language model such as BERT achieves a significantly improved accuracy in English and a moderate one in French. The most significant obstacle laying before the creation of an equally performant model in French is the lack of appropriate training data. The ideal corpus would not only contain contextually grounded lexemes, but would reflect productive rather than receptive knowledge of vocabulary.

The utility of a graded lexical classifier goes beyond Mauril's use case of vocabulary analysis. Such a model may be used in modular text simplification systems to help adjust the level of simplification and adapt it to the user's competence level. In pipelines for lexical simplification, a CEFR-based classifier might help with the ranking of substitution candidates by providing an estimation of their complexity (in context) (Gooding and Kochmar, 2019b; Aleksandrova and Dufour, 2022). It is also a fine-grained tool for complex word identification and readability analysis.

## References

Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, and Paloma Martínez. 2019. Lexical simplification approach using easy-to-read resources. *Procesamiento del Lenguaje Natural*, 63(0):95–102.

Desislava Aleksandrova and Olivier Brochu Dufour. 2022. RCML at TSAR-2022 Shared Task: Lexical simplification with modular substitution candidate ranking. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 259–263.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis.

David Alfter, Yuri Bizzoni, Anders Agebjörn, and others. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*.

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English profile wordlists project. *English Profile Journal*, 1.

Annette Capel. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

Tom Cobb. 2007. Why & how to use frequency lists to learn words.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Scott Crossley, Tom Salsbury, and Danielle McNamara. 2010. The development of polysemy and frequency use in English second language speakers. *Lang. Learn.*, 60(3):573–605.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Didier FLE. 2004. Les référentiels.

Thomas François, Mokhtar Boumedienne Billami, Núria Gala, and Dephine Bernhard. 2016. Bleu, contusion, ecchymose: tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. *JEP-TALN-RECITAL*.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. *International conference*.

Thomas Francois and Patrick Watrin. 2011. On the contribution of MWE-based features to a readability formula for French as a foreign language.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN'2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *eLex-Electronic Lexicography*.

Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Trans. Assoc. Comput. Linguist.*

Sian Gooding and Ekaterina Kochmar. 2019a. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019b. Recursive Context-Aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Gouvernement du Québec. 2014. *Agir pour interagir*. Ministère de l'Immigration, de la Diversité et de l'Inclusion.

Ray Jackendoff. 1995. *The boundaries of the lexicon. Idioms, structural and psychological perspectives*. Hillsdale, NJ: Lawrence Erlbaum.

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. Detecting multiword expression type helps lexical complexity assessment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association.

Lété. 2004. MANULEX: une base de données du lexique écrit adressé aux élèves. *Didactique du lexique*.

Deryle Lonsdale and Yvon Le Bras. 2009. *A Frequency Dictionary of French: Core Vocabulary for Learners*. Routledge.

William E Nagy. 1995. On the role of context in first- and second-language vocabulary learning. *Center for the Study of Reading Technical Report ; no. 627*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9):1–42.

Alice Pintard and Thomas François. 2020. Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92, Marseille, France. European Language Resources Association.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing knowledge-based vocabulary lists (KVL). *TESOL j.*, 12(4).

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2):251–256.