1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 85–92
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words

Alice Pintard[1], **Thomas François**[2]
[1] Le Mans Université, Le Mans, France
[2]Cental, IL&C, University of Louvain, Louvain-la-Neuve, Belgique
Alice.Pintard.Etu@univ-lemans.fr, thomas.francois@uclouvain.be

Abstract

Traditional approaches to set goals in second language (L2) vocabulary acquisition relied either on word lists that were obtained from large L1 corpora or on collective knowledge and experience of L2 experts, teachers, and examiners. Both approaches are known to offer some advantages, but also to have some limitations. In this paper, we try to combine both sources of information, namely the official reference level description for French language and the FLElex lexical database. Our aim is to train a statistical model on the French RLD that would be able to turn the distributional information from FLElex into one of the six levels of the Common European Framework of Reference for languages (CEFR). We show that such approach yields a gain of 29% in accuracy compared to the method currently used in the CEFRLex project. Besides, our experiments also offer deeper insights into the advantages and shortcomings of the two traditional sources of information (frequency vs. expert knowledge).

Keywords: Lexical difficulty, French as a foreign language, NLP

1. Introduction

Second language acquisition (SLA) research established a strong relationship between the development of reading abilities and the knowledge of vocabulary (Laufer, 1992). For Grabe (2014, 13): "The real goal for more advanced L2 reading is an L2 recognition vocabulary level anywhere above 10,000 [words]". It is no surprise that vocabulary resources used by designers of L2 curricula, publishers of educational materials, and teachers to set vocabulary learning goals are close to such size. For French language, the popular "Français Fondamental" (Gougenheim et al., 1964), which was built from a corpus of authentic documents and influenced a whole generation of French teachers and SLA researchers, includes about 8800 words. Similarly, the currently most popular lexical resources are the Reference Level Descriptions (RLDs), based on the Common European Framework of Reference for languages (CEFR), and available in various languages. The French version, designed by a team of experts, also amounts to about 9,000 words and expressions. However, both type of lists - either built from language data or from the expertise of language and teaching experts - are faced with the issue of identifying the most important words to teach at each stage of the learning process.

The most common answers to that challenge have been (1) to use frequency lists obtained from a large corpus of texts intended for native readers and split the list into N frequency bands, each of which is related to one of the stage of the learning process; or (2) to rely on expert knowledge, such as teacher expertise or linguists' recommendations, to assign each word to a given level of reading proficiency. This classification of words in developmental stages is a delicate process whose reliability has hardly been assessed in a systematic manner on L2 learners. Besides, the two main sources of information to build vocabulary lists - word frequency in massive corpora or the knowledge of L2 teaching experts - have hardly been exploited together[1].

Recently, an alternative research avenue was investigated within the framework of the CEFRLex project. It offers receptive vocabulary lists for 5 languages: English (Dürlich and François, 2018), French (François et al., 2014), Swedish (François et al., 2016), Dutch (Tack et al., 2018), and Spanish. Its innovative side resides in the fact that it does not provide a single frequency for each word, but rather a frequency distribution across the six levels of the CEFR. Moreover, frequencies have been estimated on documents intended for L2 learners, i.e. textbooks and simplified readers, instead of L1 texts. As a result, the resource provides further insights about the way a given word is used across the various development stages of the L2 curriculum. It is also possible to compare word frequency at a given level (e.g. A2) in order to define priorities in terms of vocabulary learning goals. Unfortunately, when it comes to assigning a CEFR level at which a given word should be learned, it is not obvious how the frequency distributions should be transformed in a single CEFR level.

In this paper, we aim to investigate two main issues. First, we will test whether we can leverage the knowledge from the French RLD to train a mathematical function, based on machine learning algorithms, able to transform any CEFRLex distribution into a CEFR level. Second, we will take advantage of these experiments to further characterize the linguistic and pedagogical differences between these two approaches - building a frequency list from a corpus vs. assigning words to proficiency levels based on expert knowledge - to set vocabulary learning goals. The paper is organized as follows: Section 2. provides more details about

[1]In the case of the English Vocabulary Profile (EVP), the designers have indeed combined lexicographical and pedagogical knowledge with word frequency information (Capel, 2010). However, the frequencies were estimated from a learner corpus and therefore are representative of productive skills rather than receptive ones.

the two approaches we will compare (frequency lists and RLD) and reports previous attempts to transform CEFR frequency distributions into a unique CEFR level. Section 3. introduces all methodological details related to our experiments: the three lexical resources used in the study (French RLD, Lexique3, and FLELex) and the process by which these resources were prepared for training machine learning algorithms. This section ends up with the description of the experiments carried out. In Section 4., we report the results of the various experiments before taking advantage of a manual error analysis to discuss the differences between expert knowledge and frequency-based lists at Section 5..

2. Previous work

The process of setting vocabulary goals for L2 learners generally relies on graded lexical resources in which words are assigned to one proficiency level. Such resources are usually built based on the two approaches we have previously outlined: leveraging word frequency information estimated on a corpus or using L2 teaching experts knowledge.

Frequency lists, built from a corpus, have been used since the seminal work of Thorndike (1921) who laboriously built the first significant vocabulary for English, including 20,000 words, without the help of any computer. The first computational list was obtained by Kučera and Francis (1967) from the Brown corpus and has a large influence in education and psychology. At the same time, Gougenheim et al. (1964) published the *Français Fondamental* that would impact a generation of L2 French teachers. More recently, other lists have been developed from larger corpora, such as the CELEX database (Baayen et al., 1993), the list based on the British National Corpus (Leech et al., 2001), or SUBTLEX (Brysbaert and New, 2009). The main shortcomings of such lists for L2 education are that (1) they represent the native distribution of words, which is not fully compatible with the distribution of words in books and textbooks intended for L2 learners; (2) they do not specify at which proficiency level a given word is supposed to be learned.

As regards expert knowledge, the most recent and influential resource is connected to the CEFR framework. Since 2001, this framework has been widely adopted within Europe to help standardizing L2 curricula, which involves defining a proficiency scale ranging from A1 (beginners) to C2 (mastery). However, textbook designers, assessment experts and language teachers have agreed that it lacks precision when it comes to describing the linguistic forms that should be learned at a given proficiency level. In a number of countries, efforts have been made to interpret the CEFR guidelines in the form of reference level descriptions[2]. These books describe the language competences expected from an L2 learner in each of the CEFR levels, including lists of words, syntactic structures, and expressions associated with specific communicative functions or themes.

Finally, a few papers specifically investigated methods to transform CEFRLex word distribution into a CEFR level coherent from a pedagogical perspective. (Gala et al., 2013) suggested two approaches. The first one, to which we will refer as *First occ*, assigns to a given word the level of the textbook it was first observed in. In other words, the level of a word corresponds to the first CEFR level for which FLELex reports a non null frequency. Although simplistic, this rule appeared to be the most effective to predict unknown words reported by four Dutch learners of FFL (Tack et al., 2016) and was consequently used in the CEFRLex interface. The second approach was a variant of the *First occ* that yields continuous scores and prove to be inferior to the first one. More recently, Alfter et al. (2016) introduced the concept of *significant onset of use*, that consists in selecting the first level having a sufficiently large enough delta compared to its previous level. All of these studies used mathematical rules to transform distribution into CEFR levels and later use those level as gold-standard for further process. So far, no experiments were reported that tried to cross-validate such mathematical rules, for instance using learners data.

3. Methodology

Our approach consists in considering the French RLD as a gold standard regarding the assignment of words to a given CEFR level. We then infer, from this pedagogical information, a statistical model able to transform the word frequency distribution from FLELex into a CEFR level. To carry out this experiment, the following steps had to be realized. The acquisition and digitization of the French RLD word list is described at Section 3.1., which also briefly reminds the reader of the main characteristics of the two other lexical resources used in our study, namely Lexique3 (New et al., 2007) and FLELex (François et al., 2014). In the next section (Section 3.2.), we describe a preliminary step prior to the statistical modelling, which consists in delineating the intersection between the three resources. This stage aims at ensuring that missing words would not lead to results biased towards one of the resources. We also took advantage of this step to investigate the coverage discrepancies between the French RLD and FLELex as a first way to characterize the differences between the expert knowledge approach and the frequency-based one. Section 3.3. describes the design of two datasets used for our experiments, whereas Section 3.4. presents the different baselines and models tested.

3.1. Source Word Lists

3.1.1. The French RLD word list

The RLD for French language was created by Beacco and his collaborators between 2004 and 2016 (Beacco et al., 2008; Riba, 2016). Each level – corresponding to a distinct book – is split into 10 chapters representing various dimensions of the linguistic knowledge (e.g. vocabulary, syntactic structures, phonemes, graphemes, fonctional skills, etc.), except for C1 and C2 levels which share the same volume and have a different structure[3]. The classification of linguistic forms to a given level was performed based on crite-

[2] See the list of concerned languages at `http://www.coe.int/t/dg4/linguistic/dnr_EN.asp?`

[3] The RLD book for the C levels (Riba, 2016) was not used in this study, as it doesn't provide lists of lexical items, but rather describe more conceptual abilities, like managing and structuring

Level	# FLELex	# First occ	# Beacco
A1	4,097	4,097	827
A2	5,768	2,699	615
B1	9,074	3,980	1334
B2	6,309	1,299	2742
C1	7,267	1,665	x
C2	3,932	496	x

Table 1: Distribution of entries per CEFR level, including the total number of items per level in FLELex, the number of items per level calculated with *First occ*, and the number of words per level in Beacco.

ria selected by the authors for their relevance and objectivity: essentially the official descriptors from the CEFR, collective knowledge and experience of experts, teachers and examiners, and examples of learner productions deemed to be at a particular level (Beacco et al., 2008).

To our knowledge, the French RLD, also refered to as "Beacco" in this study, has not been used so far in any NLP approaches as it was published in paper format only and is not available in a digitized version. As a consequence, we had to digitize the two chapters relative to lexicon, namely chapter 4, focusing on general notions (e.g. quantity, space), and chapter 6 that focuses on specific notions (e.g. human body, feelings, sports). Those chapters share the same architecture across all levels, organizing words within semantic categories, then specifying the part-of-speech (POS) categories and sometimes providing a context. Polysemous words can therefore have up to 8 entries across the four levels (e.g. "être", *to be*). However, as FLELex and Lexique3 do not provide fine-grained semantic distinctions for forms (all meanings are gathered under the same orthographic form), we decided to drop the information on semantic category from the French RLD. When a form had several CEFR levels associated to it, we kept the lowest one, which is in line with the way polysemy is handled in FLELex. This process led us to drop about 2,968 entries, going from 8,486 to 5,518 entries. The number of entries per CEFR level is described in Table 1 (*#Beacco*).

3.1.2. The Lexique3 word list

As previous approaches relying on word frequencies to assign proficiency levels to words relied on a L1 corpus, we decided to compare the performance obtained with FLELex with a word list whose frequencies were estimated on a large L1 corpus. We used *Lexique3* (New et al., 2007) for this purpose, as it is a rather modern database. The lexicon includes about 50,000 lemmas and 125,000 inflected forms whose frequencies were obtained from movie subtitles.

3.1.3. FLELex

FLELex (François et al., 2014) is one of the resources being part of the CEFRLex project described above. Similarly to the other languages, it offers frequency distributions for French words across the six CEFR levels. There are two versions of FLELex: one is based on the TreeTagger (FLELex-TT) and includes 14,236 entries, but no multi-word expressions as they cannot be detected by the TreeTagger; the second one is based on a conditional random field (CRF) tagger and amounts to 17,871 entries, including 2,037 multi-word expressions. However, the second version has not yet been manually checked and includes various problematic forms. This is why we decided to carry out our experiments based on the FLELex-TT version. Table 1 summarizes the total number of entries having a non null frequency per level (*#FLELex*), along with the number of new entries per level, currently used in the CEFRLex project to assign a unique level to a given word (*#First occ*).

3.2. Mapping the RLD to FLELex and Lexique3

As explained above, in order to ensure a comparability of results for each of the three word lists, we delineated their intersection. A prerequisite step was to arrange the POS tagsets compatibility. The main differences regarding those tagsets are that Beacco divides conjunctions in two categories (coordination and subordination), whereas FLELex and Lexique3 split determiners and prepositions (DET:ART vs. DET:POS and PRP vs. PRP:det). We merged all split categories, keeping the TreeTagger labels (Schmid, 1994). After standardization, nine POS remained: ADJ, ADV, KON, DET, INT, NOM, PRP, PRO, and VER.

Second, we identified words in common between Beacco and FLELex: their intersection contains 4,020 entries. This leaves 1,498 words from Beacco that do not appear in FLELex and 10,216 FLELex words absent from Beacco. Such figures were expected as the coverage of FLELex is larger due to its building principles. However, we were concerned by the fact that so many words from Beacco were not found in FLELex and carried out a manual investigation of these. Most missing words can be related to the following causes:

- Beacco includes 113 past participle forms of verbs that have not been lemmatized, whereas it is the case in FLELex (e.g. "assis" *sat*, "épicé" *seasoned*);
- Similarly, Beacco also includes 103 feminine or plural forms which are lemmatized in FLELex (e.g. "vacances" *holiday*, "lunettes" *glasses*, "serveuse" *waitress*, etc.);
- Words were sometimes shared by both resources, but were assigned a different POS-tag, preventing automatic matching (e.g. "bonjour" *hi !* or "vite" *be quick* are interjections in Beacco, but are tagged as nouns or adverbs in FLELex);
- 61 entries were kept with capital letters in Beacco as a way to provide information about the word in use (e.g. "Attention" *Look up !*, "Courage" *Cheer up !*);
- Unlike Beacco, FLELex does not include acronyms (e.g.: "CD", "DVD", "CV", etc.);
- Some words were not captured in FLELex despite their presence in FFL textbooks, because they appear in the instructions, grammatical boxes, maps, or calendars rather than in texts related to comprehension

discourse in terms of rhetorical effectiveness, natural sequencing or adherence to collaborative principles.

tasks (e.g. "fois" *time*, "adjectif" *adjective*, "virgule" *comma*, "Asie" *Asia*, etc.);

- Other words refer to very salient objects in the real world that are poorly represented in corpora. Since Michéa (1953), they are known as available words and, as was expected, some of them were not found in the corpus used to build FLELex (e.g. "cuisinière" *cooker*, "sèche-cheveux" *hair-dryer*, etc.);

- Finally, a few words in Beacco were extremely specific (e.g. "humagne", a type of grape or "escrimeur" *fencer*).

This manual investigation was, to some extent, reassuring, as a fair amount of missing words from Beacco were due to discrepancies in the lemmatization process between a systematic tool and a human. Lexical availability was also an issue, but a predictable one as it concerns all frequency-based approaches. Finally, it appears that restricting the selection of textbook materials to texts related to receptive tasks might help to better model receptive knowledge of L2 learners, but also comes at a cost as regards coverage.
We manually solved some of these issues by lemmatizing the entries; converting the POS of all interjections that were nouns or adverbs in FLELex, and replacing capital letters by lowercase letters. In this process, we lost precious information from the RLD about the function of some linguistic forms, but were able to reintroduce 314 words that were not considered as shared by both lexicons before. As a result, the intersection between both resources amounts to 4,334 words. Finally, in order to compare FLELex with Lexique3, we computed the intersection between all three lexicons. Lexique3 having the larger coverage (51,100), there were only 38 words missing from it. The final intersection therefore includes 4,296 entries.

3.3. Preparing datasets for experiments

Based on this intersection between the three resources, we defined two datasets that will be used for our experiments.

3.3.1. BeaccoFLELexAtoB

This first dataset corresponds to the intersection between FLELex, Lexique3 and Beacco as defined at Section 3.2.. It contains 4,296 entries, shared by the three lexicons, and classified from A1 to B2 according to Beacco. In this dataset, each entry (word + POS-tag) is related to its CEFR reference level from Beacco and is described with 8 frequency variables, as shown in Table 2. The frequency variables includes the 7 frequencies provided by FLELex along with the frequency from Lexique3. The latter will however be used only for the computation of the Lexique3 baseline (see Section 4.).

3.3.2. BeaccoFLELexC

The main application of this study is to develop a more accurate mathematical model to transform FLELex frequencies into a single CEFR level, with the purpose of integrating this model within the web interface of the CEFRLex project instead of the *First occ* heuristic currently used. Therefore, training our model on the intersection described above has a main shortcoming: it is not able to classify any entries beyond B2 level, since it would not have seen any word from the C levels. In the FLELex interface, we nevertheless want to be able to classify words at those levels, as FLELex likely contains more difficult words than Beacco.
To create this second dataset (*BeaccoFLELexC*), we first assumed that the 9,903 FLELex entries missing from Beacco can be considered as C level. However, before adding these entries to the 4,296 word intersection, we manually investigated them and noticed that about 2% present high frequencies in A levels textbooks, which is not expected for C words. We thus considered these cases as anomalies. Some causes of these anomalies were already discussed previously, but new issues also arose:

- Function words appearing in Beacco's chapter 5, i.e. the grammar section, were not digitized, but they were logically captured in FLELex. They include personal pronouns ("je", "tu", "toi"), interrogative pronouns ("combien", "où", "comment", "quand"), determiners ("la"), prepositions ("en", "sur"), conjunctions ("après", "pour", "que"), modals ("devoir", "pouvoir"), and negative particles ("non", "ne", "pas");

- We also identified a few words appearing in chapter 3, linked to particular communicative functions, that were also excluded from our digitizing process (e.g. "cher" *dear*, "bise" *kiss*, "peut-être" *maybe*, "d'accord" *all right*, etc.);

- Other words are very likely part of the A levels even if they are not included in Beacco's chapters we digitized (e.g. "joli" *pretty*, "dormir" *to sleep*, "anglais" *English*, or "espagnol" *Spanish*);

- Finally, we identified a few remaining tagging problems in FLELex that escaped the manual cleaning process (e.g. "étudiant" *student*, "ami" *friend* were found as adjectives in FLELex instead of nouns).

To resolve some of these issues, we manually corrected tagging problems in FLELex and added the missing words appearing in chapters 3 and 5, assigning them their correct Beacco level. In total, 87 words were thus corrected, but some problems remain for a few entries.
The last step in the preparation of this dataset *BeaccoFLELexC* consisted in creating a balanced dataset. Adding 9,903 C entries obviously produced a class-imbalanced issue within the data, which we rebalanced using undersampling of overrepresented categories (C and B2). We used a random undersampling technique based on the number of entries in B1, reducing the size of this dataset from 14,236 to 4,878 words.

3.4. Experiments

For our experiments, we decided to use three standard machine learning algorithms, namely tree classification, boosting, and support vector machine (SVM). Neural networks were not considered due to the limited amount of data. We also defined four baselines to compare with, that are described below.
All experiments were conducted following the same methodology. We first split each dataset into a training

word	pos	beacco	freqA1	freqA2	freqB1	freqB2	freqC1	freqC2	freqtotal	lex3
plier	VER	B2	0.00	2.14	5.15	13.73	3.44	12.83	8.37	14.37
chanteur	NOM	A2	46.75	42.96	21.32	18.26	3.44	50.42	36.12	21.17
humide	ADJ	A1	0.00	0.00	13.94	0.00	18.00	0.00	5.36	11.23
entre	PRP	B1	601.22	995.40	1023.06	774.83	1599.32	2023.56	1032.37	372.72

Table 2: Examples of entries for "plier" *to fold*, "chanteur" *singer*, "humide" *humid* and "entre" *between* from the first dataset, illustrating the variables used in our experiments.

(and validation) set including 80% of the entries and a test set including 20% of the entries. We then applied a grid search on the training set using a stratified 10-fold cross-validation setup to estimate the performance of each set of meta-parameters tested. Once the best set of meta-parameters was chosen, we estimated the classification accuracy of the model on the test set. This procedure is more reliable than a standard 10-fold cross-validation setup as the meta-parameters and the parameters are not optimized on the same data.

3.4.1. Baselines

Four baselines were used in this study. The first one (*Maj class*) assigns to all words the level of the majority class. It is a common baseline for all classification tasks. The second baseline (*First occ*) assigns to a given word the level of the textbook it was first observed in. The third baseline (*Most freq*), used for instance in Todirascu et al. (Todirascu et al., 2019), assigns to each word the level with the highest frequency. For the fourth baseline, we trained three models (SVM, tree, and boosting) based only on Lexique3 frequencies, as a way to assess whether the L2-specific and more fine-grained frequency information from FLELex would lead to some improvements on the task.

3.4.2. The models

We applied the three above-mentioned algorithms to both our datasets: *BeaccoFLELexAtoB* and *BeaccoFLELexC*.

- On the former, the optimal meta-parameters found by the grid search for Lexique 3 were: Tree (max_depth = 4, min_sample_leaf = 40, and min_sample_split = 50); SVM (RBF kernel with C = 0.01 and γ = 0.001); Boosting with 5 iterations.
- The meta-parameters found for FLELex frequencies were: Tree (max_depth = 3, min_sample_leaf = 20, and min_sample_split = 50); SVM (RBF kernel with C = 1 and γ = 0.0001); Boosting with 5 iterations.
- On the latter, *BeaccoFLELexC*, the optimal meta-parameters found using the grid search were: Tree (max_depth = 3, min_sample_leaf = 20, and min_sample_split = 50); SVM (RBF kernel with C = 1 and γ = 0.001); Boosting with 5 iterations.

4. Results

In this study, we aim to predict L2 expert knowledge based on word frequencies and thus obtain a machine learning algorithm able to transform a given word's frequency into a unique CEFR level. First, our systematic evaluation on the *BeaccoFLELexAtoB* dataset, whose results are reported in Table 3, reveals that the *First occ* rule, currently used in the CEFRLex interface, yields poor performance. Its accuracy is as low as 25%, which is actually lower than the accuracy reached by a majority class classifier (40%) and its mean absolute error is 1.25, which means that this classification rule can miss targeted levels by even more than one level on average. Similarly, the *Most freq* rule, sometimes used as a simple and intuitive solution by some researchers, appears to be quite disappointing: its accuracy of 18% reveals that it is actually biased towards wrong answers. Using a machine learning algorithm to train a non-linear and more complex mathematical rule to transform FLELex distributions into CEFR levels seems to be a better path. We were able to reach 54% for the Boosting classifier and a mean absolute error of 0.66. The SVM model is more than twice as good as the *First occ* rule, and it outperforms the majority class classifier by 13%.

BeaccoFLELexAtoB				
	Acc	Prec	F1	MAE
First occ	0.25	0.45	0.21	1.25
Most freq	0.18	0.35	0.23	1.62
Maj class	0.40	0.16	0.23	1.13
Lexique3 frequency				
Tree	0.47	0.46	0.46	0.76
SVM	0.49	0.44	0.43	0.76
Boosting	0.49	0.38	0.39	0.80
FLELex frequencies				
Tree	0.52	0.52	0.52	0.68
SVM	0.53	0.48	0.46	0.68
Boosting	0.54	0.51	0.48	0.66
BeaccoFLELexC				
	Acc	Prec	F1	MAE
First occ	0.27	0.33	0.23	1.35
Most freq	0.19	0.23	0.20	1.69
Maj class	0.22	0.05	0.08	1.19
Tree	0.47	0.46	0.46	0.76
SVM	0.44	0.41	0.40	0.87
Boosting	0.48	0.45	0.45	0.75

Table 3: Test results on both datasets.

On the second dataset, that corresponds better to the pragmatic problem we want to solve, it is interesting to notice that *First occ* outperforms the dummy baseline using majority class by 5%. The *Most freq* rule remains the worst option, whereas machine learning remains the best with the boosting algorithm reaching 48% of accuracy and a MAE of 0.75. Performance are slightly behind for the second dataset, but this is generally the case when one increases the number of classes.

We also performed an ablation study on both datasets in order to find which frequency level contributed the most to the predictions. Results are presented in Table 4 and clearly shows that the frequency from the A1 to B1 levels are the more informative, especially the A1 level. Furthermore, one can notice that the total frequency (computed over all six levels) is also a good source of information.

4.1. FLELex vs. Lexique 3

In our experiments, we wanted to know whether the L2-specific and fine-grained frequency information provided in the CEFRLex resources would be better able to predict expert knowledge than a L1 frequency list. Table 3 shows that the models trained with FLELex slightly outperform (+5% in accuracy) the ones trained with Lexique3. However, this comparison is unfair, as the models leveraging FLELex information include more variables than the Lexique3 ones (7 vs. 1). Looking at the ablation study table, we can see performance when only the total frequency variable of FLELex is used. In such configuration, FLELex still outperforms Lexique3 by 1% accuracy, which seems to mean that L2 frequencies - even estimated on a much smaller corpus - might be used instead of L1 frequencies. This is, per se, a very interesting result, as the second language acquisition literature tends to believe the opposite and L2 intended word list created from a L1 corpus still remains the standard. In any case, those similar results can also be explained by the high correlation between the frequencies of these two lists, as was already reported in François et al. (2014). If we consider the performance of the full model (54%) compared to that of the model based only on the total frequency, the 4% improvement could be interpreted as a confirmation of the greater informational richness provided by a frequency distribution over proficiency levels compared to a unique word frequency.

	BeaccoFLELexAtoB		**BeaccoFLELexC**	
Variable	All but 1	Only 1	All but 1	Only 1
freqA1	0.54	0.53	0.43	0.43
freqA2	0.53	0.52	0.47	0.40
freqB1	0.54	0.52	0.47	0.40
freqB2	0.54	0.47	0.47	0.39
freqC1	0.54	0.45	0.47	0.36
freqC2	0.55	0.43	0.47	0.34
freqTotal	0.54	0.50	0.46	0.44

Table 4: Variable ablation study on both datasets, using the boosting model.

4.2. Problematic levels

The analysis of the precision, recall, and F1 values for each level reveals that models predictions are affected by one level in particular, level A2, which is already underrepresented in Beacco. Hence, the *BeaccoFLELexAtoB* dataset only includes 573 words at this level, whereas A1, B1 and B2 levels contain respectively 788, 1158 and 1777 words. Table 5 shows that extreme levels score much better than the middle ones, a recurrent outcome in readability classification tasks. It also reveals that, despite its lower accuracy score compare to the boosting model, the classification Tree model takes less drastic choices when assigning words to a class, which makes it a better option if we want a system that assigns words to all levels. We also noticed that, besides their under-representation in the RLD, A2 words are difficult to predict due to a high correlation between word frequencies in A1, A2 and B1 levels.

Level	**Tree**	**SVM**	**Boosting**
A1	0.57	0.56	0.58
A2	0.21	0.13	0.02
B1	0.26	0.23	0.32
B2	0.67	0.69	0.70

Table 5: F1 scores per level for the three models, on the *BeaccoFLELexAtoB* dataset.

Another problematic level is the C level, specially from a reading comprehension perspective. According to the CEFR descriptors, a C1 user "can understand a wide range of demanding, longer texts, and recognise implicit meaning", while a C2 user "can understand with ease virtually everything heard or read". Trying to translate these descriptors into actual words is difficult, as testified by the fact that Riba, who wrote the RLD opus for C levels, expressed some reserves concerning those descriptors, mainly because of the notion of perfection which emanate from them (Riba, 2016), and the fact that C users depicted by the CEFR are only highly educated individuals, outperforming most of the native speakers. Consequently, we had to use a simple decision to define our C words in the gold-standard: considering everything minus words from levels A1 to B2. A final issue regarding C words is the fact that textbooks for those levels are less numerous than for the lower ones, providing FLELex with fewer words to observe and count.

5. Discussion

In this section, we carried out a manual error analysis of some misclassification errors as a way to bring up to light some strengths and weaknesses of both approaches that can be used for selecting appropriate vocabulary in L2 learning: frequency vs. expert knowledge.

5.1. Lexical approach

One characteristic of the RLDs that is worth remembering is the fact that lexical chapters are organised semantically, as the authors agreed that giving a list of words ranked alphabetically is of little use when it comes to design syllabus or build a teaching sequence (Beacco et al., 2008). Hence, words evolving around the same notional scope come together, along with their synonyms, antonyms and as a matter of fact words belonging to the same family as well (e.g. "heureux/malheureux" *happy/unhappy*, "maigre, gros/ maigrir, grossir" *skinny, fat / to loose weight, to put on weight*). This conveys the idea that they should be taught together - in other words, at the same CEFR level – since building strong lexical networks is critical for vocabulary retention. Conversely, FLELex does not have such structure and is likely to estimate different frequency distributions

for the various terms from a given semantic field. When we transform those distributions using either the *First occ* rule or even a machine learning algorithm, they are prone to end up at different levels (e.g. of predictions using the SVM: "gros"A2 / "grossir"B2, "heureux"A1 / "malheureux"A2). In this regard, using frequency lists to establish a vocabulary progression through several learning stages is limited because words are seen as isolated.

Beacco's semantic organisation also enables it to better capture the effect of situation frequency, usually referred to as lexical availability (Michéa, 1953). The B2 level was the first RLD written, and it consists of extensive lists of words relating to specific centers of interest. The lower levels were compiled later, gradually picking up words from the B2 RLD according to the learning stage they could be taught at. As a result of this semantically-driven procedure, Beacco includes more available words than FLELex (e.g. of missing available words are "soutien-gorge" *bra*, "cuisinière" *cooker*, "sèche-cheveux" *hair-dryer*, etc.).

5.2. Topics

The question of topics covered on the L2 learning path is very relevant for this study, because it highlights the limits of both methods. FLELex computational approach aims to report words frequencies in the CEFR levels in an objective and descriptive way, but by using L2 material, it is compelled to favour certain topics and exclude others. Compared to Beacco, we found that textbooks mainly avoid potentially polemic themes such as religion or death, or subjects in which students could have not much to say such as DIY, and topics commonly considered as complicated, for instance economics or sciences. In contrast, the topics found in FLELex are highly influenced by the choice of texts comprised in the corpus and can sometimes be overrepresented. A clear materialization of this shortcoming appeared when we looked at FLELex frequent words absent from Beacco and discovered that words related to fairytales were abundant (e.g. "château" *castle*, "reine" *queen*, "prince" *prince*, "chevalier" *knigth*, "dragon" *dragon*, "magique" *magic*, and even "épouser" *to marry* or "rêver" *dream*). This can be explained by the inclusion of a simplified reader dedicated to King Arthur legend in the corpus.

On the other hand, the RLD's semantic structure has a downside since it may lead to loose sight of the CEFR descriptors, specially in topics where finding a progression between the items in terms of language acts is arduous. The most iconic theme we came across is food and drinks, with 150 words absent from FLELex, but geography and the human body also share the same characteristics at a lesser degree. We distinguished those topics from the others because they are mostly composed of nouns with closely related meaning (e.g. in B2, "pain français", "baguette", "boule", "bâtard", "pain de campagne", "carré", "pain intégral", "pain complet", "pistolet", "petit pain", "sandwich", "petit pain au lait", all being different types of bread). The large number of words in these topics is a reflection of reality usually bypassed in textbooks, since these nouns don't offer a wide variety of communicative situations.

5.3. Reception or production

FLELex is a descriptive tool built from texts related to reading comprehension tasks in FFL materials, illustrating therefore the contents of written reception activities. The RLD also presents its contents as to be mastered in comprehension tasks, leaving the decision to teachers and curriculum designers regarding what learners should be able to produce (Beacco et al., 2008). However, we identified four POS in which the ability to produce words seems to be the selection criteria for Beacco: determiners, conjunctions (e.g. "comme" in B1), pronouns (e.g. "le" in B2), and prepositions. We detected them because the frequencies of those POS are among the highest of the corpus while their levels nevertheless vary from A1 to B2 in Beacco. Even though words belonging to these POS are probably understood at early stages due to repeated exposure, the RLD proposes a gradation in the different learning stages they should be taught at, which is likely motivated either by the CEFR descriptors regarding production and interaction or by intrinsic characteristics of the word. We therefore found that the two approaches are not compatible for those specific POS, as the prescriptive aspect of the RLD implies to take into account learners objectives and abilities in production tasks as well, while FLELex only illustrates the language used in reception tasks.

5.4. Normative and adaptable

Beacco's intent is to propose a reference calibration for CEFR levels, but not a list of words that would be mandatory and identical, in all places and at all times. In the introduction, the authors minimize the inherent normative aspect of their lists, presenting them as only a starting point to insure compatibility between syllabus and exams of different educational systems. Therefore, they display vocabulary in three possible ways:

- closed lists, e.g. "bébé, enfant, lait"
- open lists, e.g. "[...] agréable, bête, calme, content"
- list descriptors, e.g. "[...] *noms de nationalités*"

Such behavior, intended to human readers, however raises some issues for an automatic approach. Facing list descriptors, we generally ignored them in the digitizing process, which explains why words such as "anglais" *English* and "espagnol" *Spanish* – which are nationalities – were not found in our version of Beacco, although present in FLELex. For our study, open lists and list descriptors are very problematic in the sense that the absence of a word from a level cannot be considered as 100% certain. From a teacher's perspective though, those open lists and item descriptions are coherent with the authors goal to provide content adaptable to all contexts, and indications that the items are to be chosen according to the geographic, cultural and educational situation (e.g. for the nationalities, "japonais", "coréen" and "vietnamien" are likely to be taught in A1 to Asian learners, whereas they might not be needed from A1 in a South American classroom).

6. Conclusion

In this research, we aimed to infer CEFR levels from CEFRLex word frequency distribution using expert knowledge

from the French RLD as a gold-standard. Such approach enabled us to apply, for the first time, machine learning algorithms to such task whereas previous work used simple mathematical rules. After standardisation of the data, we trained three machine learning models on two sets, reaching an accuracy score of 0.54 for the dataset *BeaccoFLELexAtoB* and of 0.48 for the *BeaccoFLELexC* dataset. These results clearly outperforms results reached by the *First occ* rule currently used in the CEFRLex interface. Our work has direct repercussions on this project, as our best classifier has been integrated in the interface[4], offering now the choice between *Beacco* or *First occ* to classify words.

Our experiments also yield other interesting results. First, comparing our results with those of a L1 frequency word list revealed that the distributional information contained in FLELex indeed seems richer and finer-grained than the one of a standard L1 list. Second, we carried out an analysis on the most important classification errors as a way to sharpen our understanding of the differences existing between the two approaches we compared: frequency and expert knowledge. This analysis stressed the importance of lexical networks in L2 learning to ensure a better representation of available words and of words connected to topics generally avoided in textbooks. We also noticed that although CEFRLex resources only represent receptive skills, Beacco might have sometimes classified words based on criteria relative to both receptive and productive skills. Finally, the presence of list descriptors in RLD is a serious issue for their automatic exploitation, as they contain some implicit knowledge. We believe that all these discrepancies partially explain why our statistical model is not able to better predict Beacco's level. In other words, although a better option than the *First occ* rule, using expert knowledge also has shortcomings. In the future, we plan to investigate the use of L2 learners data as an alternative source of information to transform CEFRLex distribution into levels.

7. Bibliographical References

Alfter, D., Bizzoni, Y., Agebjórn, A., Volodina, E., and Pilán, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP4CALL and NLP for Language Acquisition at SLTC*, number 130, pages 1–7. Linköping University Electronic Press.

Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). *The {CELEX} lexical data base on {CD-ROM}*. Linguistic Data Consortium, Philadelphia: Univ. of Pennsylvania.

Beacco, J.-C., Lepage, S., Porquier, R., and Riba, P. (2008). *Niveau A2 pour le français: Un référentiel*. Didier.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Capel, A. (2010). A1-b2 vocabulary: Insights and issues arising from the english profile wordlists project. *English Profile Journal*, 1(1):1–11.

Dürlich, L. and François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of LREC 2018*, pages 873–879.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of LREC 2014*, pages 3766–3773.

François, T., Volodina, E., Ildikó, P., and Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of LREC 2016*, pages 213–219.

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 *Proceedings of eLex2013*, pages 132–151.

Gougenheim, G., Michéa, R., Rivenc, P., and Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré)*. Didier, Paris.

Grabe, W. (2014). Key issues in l2 reading development. In *CELC Symposium Bridging Research and Pedagogy*, pages 8–18.

Kučera, H. and Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In *Vocabulary and applied linguistics*, pages 126–132. Springer.

Leech, G., Rayson, P., and Wilson, A. (2001). Word frequencies in written and spoken english: based on the british national corpus.

Michéa, R. (1953). Mots fréquents et mots disponibles. un aspect nouveau de la statistique du langage. *Les langues modernes*, 47(4):338–344.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677.

Riba, P. (2016). *Niveaux C1 / C2 pour le français: éléments pour un référentiel*. Didier.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016). Evaluating lexical simplification and vocabulary knowledge for learners of french: possibilities of using the flelex resource. In *Proceedings of LREC2016*, pages 230–236.

Tack, A., François, T., Desmet, P., and Fairon, C. (2018). NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of BEA 2018*.

Thorndike, E. (1921). Word knowledge in the elementary school. *The Teachers College Record*, 22(4):334–370.

Todirascu, A., Cargill, M., and François, T. (2019). Polylexfle: une base de données d'expressions polylexicales pour le fle. In *Actes de la conférence TALN 2019*, pages 143–156.

[4]The interface is available at `https://cental.uclouvain.be/cefrlex/analyze`.