FLELex: a graded lexical resource for French foreign learners

T. François (1), N. Gala (2), P. Watrin (3), C. Fairon (1)
(1) CENTAL & ILC, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
(2) LIF-CNRS & Aix Marseille Université, 13288 Marseille, France
(3) EarlyTracks, Louvain-la-Neuve, Belgique

thomas.francois@uclouvain.be, nuria.gala@lif.univ-mrs.fr

Abstract

In this paper we present FLELex, the first graded lexicon for French as a foreign language (FFL) that reports word frequencies by difficulty level (according to the CEFR scale). It has been obtained from a tagged corpus of 777,000 words from available textbooks and simplified readers intended for FFL learners. Our goal is to freely provide this resource to the community to be used for a variety of purposes going from the assessment of the lexical difficulty of a text, to the selection of simpler words within text simplification systems, and also as a dictionary in assistive tools for writing.

Keywords: graded lexicon, CALL, FFL vocabulary learning

1. Introduction

Language technologies offer new possibilities for vocabulary learning and for writing (i.e. assistive technologies for comprehension and production tasks). In computer-assisted language learning (CALL) applications (e.g. for French foreign learners (FFL): *ALFALEX* (Selva et al., 2004), *COBRA* (Deville et al., 2013), among others), lexical resources are usually offered to learners, but these resources are tailored for humans and could hardly be used as it for developing NLP applications. Moreover, electronic dictionaries and lexical resources from learning platforms lack of explicit information on the levels of difficulty of words. For all these reasons, such resources are not appropriate for being used in natural language processing (NLP) applications, such as automatic text simplification or the assessment of text readability.

As far as second language acquisition is concerned, to our knowledge, the only available resource that classifies words in various levels are the CEFR referentials (for French, (Beacco and Porquier, 2007) (A1), (Beacco et al., 2007) (A2), etc.). The CEFR scale (*Common European Framework of Reference for Languages*) (Conseil de l'Europe, 2001) defines six levels of proficiency that ranges from A1 (basic knowledge) to C2 (proficiency) and provides educative guidelines for professionals in second language teaching. Again, for NLP purposes, those materials have several shortcomings. First, words are organised in structured themes, defined by CEFR experts, and selected according criteria that might not be corpus-based. Second, a word can be listed in several levels and no further discrimination is done regarding its relative importance across the levels. Finally, and even more problematic, there is no electronic version of such corpora, which considerably hinder its use for NLP tasks such as text simplification.

For all these reasons, we propose to build a graded lexicon better describing the behavior of words across the CEFR levels. The paper is organized as follows: we first provide a background on lexical resources for language learning. In section 3, we describe the methodology applied to build our FFL graded lexicon: collecting data from textbooks, scanning and OCR, tagging and finally computing different formulae (raw frequencies, dispersion, etc.). Section 4 presents the resource and discuss the data. To conclude, section 5 provides an overview and explores improvements and applications.

2. Related work

Resources for learning languages and psycholinguistic studies have significantly changed since the incorporation of frequency values and, more recently, with the exploitation of very large annotated corpora.

2.1. Lexical lists for vocabulary learning

From the very beginning of lexicography, creating wordlists was mainly motivated by pragmatical purposes, such as providing teachers with the words that should be instructed in priority. The first lists to include quantitative data started to appear from the early 20th century. The *The teacher's word book* of (Thorndike, 1921) (for English) is one of the most famous. It is a list of 10,000 words ranked according to their frequency of occurrence in a corpus of 4,500,000 words sampled from children books, technical textbooks, newspapers articles, etc. Thorndike helped to lay the foundations of the use of statistical data for pedagogical purposes, being one of the first to argue that the more frequent a word is, the more adequate it is for young readers.

Numerous studies stemmed from this seminal work. The Thorndike's list was used in several readability formulas to help measuring the reading difficulty of texts (Lively and Pressey, 1923; Vogel and Washburne, 1928). Other similar lists flourished for other languages: the *French Word Book* of Henmon (1924), the *Spanish Word Book* of Buchanan (1927), and the *French Word Book* of Vander Beke (1932). Thorndike's list itself was also expanded some years later to 30,000 words by Thorndike and Lorge (1944).

All these resources are based on the assumption that the word frequency effect is a good predictor of word recognition performances. The word frequency effect has been

mentioned first by Cattell (1885), then experimentally confirmed by Howes and Salomon (1951) as well as by more recent research (Monsell, 1991; Brysbaert et al., 2000). The explanation for this effect seems to be that "the representations of common words in the mental lexicon are more easily accessed than those of less common words (e.g., due to a lower threshold or to an elevated activation level)" (Brysbaert et al., 2000, 66). Moreover, this effect impact mostly the decoding phase of the reading process (i.e. the step in which words are identified), since (Solomon and Postman, 1952) found an effect even for words whose meaning was unknown from the subjects. However, it is agreed that better decoding skills support the comprehension step, since less mental resources are required to perform the decoding, leaving more resources available for the comprehension processes.

Subsequently, several shortcomings of this frequentist approach of the lexicon were raised. First, words must be seen a sufficient amount of times to get a robust estimation of their frequency. Thorndike (1921) already reported that the values obtained for the first half of his list were more robust than those from the second half, even tough his corpus was quite large for the time being. Second, some words are common in the language (such as *toothpaste*, *miniskirt* or *ceiling*), but are rarely attested in written texts, the documents generally used for frequency estimation. This type of words were called available words by Michéa (1953) who took part in the elaboration of one of the most important pedagogical list for French: the *Dictionnaire fondamental de la langue française* by (Gougenheim, 1958). Gougenheim list was intended to help people learn French as a foreign language. They contain basic French words, selected both based on frequencies in a corpus and among the most salient available words. For French, we can also mention the *Listes orthographiques de base du français* by (Catach, 1984), which was created to help schoolchildren to spell French words correctly.

2.2. Computational resources with quantitative information

With the development of corpus linguistics and computational linguistics, the quantitative approach of the lexicon expanded (e.g. works on lexical statistics such as those of (Church and Hanks, 1990) and (Church et al., 1991) among others). It was then possible to gather large corpora and automatically compute frequencies. Based on the *Brown Corpus*, Francis and Kucera (1967) thus defined a new frequency list for the words in American English. Using a balanced corpus, they noticed that frequency distributions depend on the type of documents used in the corpora as well as on the topic covered in those documents. If a word is frequently used in a few number of texts because it is related to the topic, the frequency of this word could be overestimated. To prevent this limitation, more complex frequency counts were considered, such as the dispersion, the standard frequency index, etc. In subsequent lists, distributional properties of words (collocations, n-grams, etc.) were also considered (e.g. the *British National Corpus* (BCN) (BNC-Consortium, 2001)). The linguistic information in these resources was also enhanced with the addition of part-of-speech tags, phonological patterns, etc.

Machine-readable corpora were also used for the constitution of lexical databases intended for psycholinguistic studies, i.e. research on the reading processes or the language acquisition. Brulex (Content et al., 1990) is the first resource of this type describing the French language. More recently, Lexique3 (New et al., 2001) reports linguistic and frequential information for 47,342 lemmas and it has been used both for psycholinguistic studies and for natural language processing (NLP) research. Last but not least, the *French Lexicon Project* is a resource used in lexical decision tasks. It involved the collection of 38,840 French words and the same number of non-words across 1,000 participants from different French universities (Ferrand et al., 2010) (inspired from a similar project for English (Balota et al., 2007)).

By and large, such resources are relevant for a psycholinguistic analysis of the reading processes in adults, as well as for NLP tasks assuming a standard view of the language. However, they lack information about how words are used by populations having a different level of knowledge of the language, such as children learning their mother-tongue or foreign language learners.

2.3. Graded resources

Information on the difficulty of the vocabulary may be very useful in a variety of domains such as language learning, readability assessment, or automatic text simplification. However, except for scholar dictionaries with 'simple' words intended for children learning their mother tongue (e.g. *The American Heritage Student Dictionary*, the *Larousse des débutants*, etc.), dictionaries with information on the levels of the difficulty of the words are extremely rare.

To our knowledge, the only graded-lexicon available for French is Manulex (Lété et al., 2004). This database contains frequencies accounting for the presence of a word in a particular grade of elementary school textbooks (1st grade, 2nd grade and higher grades). More recently, (Gala et al., 2013) developed ReSyf, a graded lexicon of synonyms compliant with those three Manulex levels, using a SVM model to predict lexical difficulty for unseen words. The predictions are based on a set of linguistic and psycholinguistic features gathered from different lexical resources.

However, as mentioned previously, such lexical resources do not exist for French as foreign language, although it is a domain for which being able to relate lexical forms with levels of proficiency is a crucial task. In this paper, we present a graded-lexicon inspired from Manulex, but intended to learners of French as a foreign language (FFL) and compliant with the CEFR levels of proficiency. By graded, we mean that each word is presented along with its frequency distribution computed across the CEFR levels. The next section details the methodology used to obtain such a resource.

3. Methodology

Our FFL lexicon is intended both for NLP tasks and language learning purposes, which entails that word distributions have to be computed on text which are representative

of the documents used for teaching. Furthermore, these texts have to be classified according to a widely-spread scale of proficiency. Section 3.1. explains how we settle these two issues and presents the corpus used to estimate the frequency distributions for every word. In section 3.2., we describe the part-of-speech tagging required to yield a resource in the form of a list of lemmas along with their POS. Finally, section 3.3. introduces the formulae applied to the raw frequencies to get better predictors of the actual frequency distribution of words.

3.1. Source corpora

We collected a large number of texts that were classified according to a widely-spread scale of proficiency. As already mentioned above, the obvious choice was the CEFR scale that comprises the six following levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). It has indeed become the reference for second language teaching within Europe. However, to find a large number of texts following this scale is not an easy task. To our knowledge, there is no digital resource freely available that contains a large amount of texts for FFL annotated in terms of the CEFR levels. To build our graded-lexicon for FFL, we thus had to manually collect texts from printed textbooks and simplified readers that were compliant with the CEFR scale.

Among all available textbooks, 28 textbooks and 29 readers were selected depending on the two following criteria: (1) they had to be published after 2001 and (2) they must be intended for adults or teenagers learning FFL for general purposes. With these criteria, we extracted 2,071 texts related to a reading comprehension task and we assigned to each of them the same level as the textbook or reader it came from. Afterwards, all texts were scanned and automatically transformed into a machine-readable format (XML). To perform this task, we used optical character recognition tools and we manually revised and corrected the texts. The resulting corpus includes about 777,000 words, distributed across several textual genres or types as described in Table 1.

The category *Varias* includes documents such as ads, songs, poems, recipes, etc. while the category *Texts* includes texts from textbooks that are mostly informative texts along with some narrative ones. The category *Readers* comprises all texts from the simplified readers, that are longer and more coherent than textbook documents. It should also be mentioned that although the corpus does not seem very balanced across text genres and levels at first glance, we believe that these figures are pretty representative of the distribution of texts within the FFL textbooks of our population.

3.2. Tagging the data

Once the corpus gathered, the next step was to tag every texts. We wanted to obtain the lemma of every form observed in the corpus and to disambiguate homographic forms with different part-of-speech tags (e.g. *général* which can be a noun or an adjective). Inflected forms could also have been considered, but this entails that words having numerous inflected forms, such as verbs, would have their overall probability split between their different forms. Consequently, compared to invariable words (such as adverbs, prepositions, conjunctions), they would seem less frequent than they really are. Second, using tokens presupposes the assumption that learners are not able to relate inflected forms with their lemma. Such a view seems highly questionable for most of the French words that have regular inflected forms.

Another issue we faced was the detection of multi-word expressions (MWEs) in the texts. The class of MWEs gathers a set of heterogeneous linguistic objects, the meaning and structure of which can be more or less frozen (collocations, compound words, idioms, etc.). From a statistical point of view, this class of objects commonly refers to "strings of words that are more frequently associated than it would be only by chance" (Dias et al., 2000, 213). For L2 learners, it has been demonstrated that their MWEs knowledge lags far behind their general vocabulary knowledge (Bahns and Eldaw, 1993). Therefore, including such linguistic forms in a graded-lexicon for FFL purposes appears as a requirement. The tagger we first considered, TreeTagger (Schmid, 1994), is a well-known and widely-used tagger within the NLP community. However, its accuracy on real texts is now behind current state-of-the-art taggers. In addition, its major drawback is not to be able to detect multiword expressions. We thus applied a second tagger to our corpus, based on the work of Constant and Sigogne (2011). This tagger combine a conditional random fields model and large coverage linguistic resources (including MWE). This tagger reaches higher performance than TreeTagger on newspaper articles and is also able to detect some MWEs (its efficiency on narrative texts, poems or dialogues remains nevertheless unreported).

To create our FFL lexicon we took into account the performances of both taggers, as tokenization and tagging are crucial in a lexical resource. Errors at this stage produce unwanted effects on the data such as:

- entries with wrong part-of-speech tag (e.g. *adoptez* 'you adopt' PREP, *tu* 'you' ADV);
- entries with a non attested lemma (e.g. *faire partir* 'drive someone away' instead of *faire partie* 'to be part of', *peux* instead of *pouvoir* 'to can');
- tags that are likely to be, but are erroneous in the specific context of the word (e.g. to tag as an adverb the word *forward* in *the forward part of the ship*)[1].

A manual validation could have been useful to remove the two first kind of errors. However, this will also lead to a loss of the probability mass. As a consequence, we decided to assess the performance of both taggers used in this study to get an idea of the confidence that can be granted to the frequency estimation process. Although both taggers have already been assessed elsewhere, we wanted to get an estimate of their efficiency on our specific corpus.

[1]This type of error does not lead to the creation of a wrong entry, but mess up the frequency estimations, since the word occurrence will be assigned to the wrong entry.

Genre	A1	A2	B1	B2	C1	C2	Total
Dialogue	153 (23,276)	72 (17,990)	39 (11,140)	5 (1,698)	/	/	269 (54,104)
E-mail, mail	41 (4,547)	24 (2,868)	44 (11,193)	18 (4,193)	8 (2,144)	1 (398)	136 (25,343)
Sentences	56 (7,072)	21 (4,130)	12 (1,913)	5 (928)	/	/	94 (14,043)
Varias	31 (3,990)	36 (4,439)	23 (5,124)	14 (1,868)	1 (272)	/	105 (15,693)
Text	171 (23,707)	325 (65,690)	563 (147,603)	156 (63,014)	175 (89,911)	48 (34,084)	1,438 (424,009)
Readers	8 (41,018)	9 (71,563)	7 (73,011)	5 (59,051)	/	/	29 (244,643)
Total	460 (103,610)	487 (166,680)	688 (249,984)	203 (130,752)	184 (92,327)	49 (34,482)	2,071 (777,835)

Table 1: Number of texts and words per text category in the corpus

The evaluation process was carried as follows. First, one hundred sentences were sampled from the corpus and tagged with both taggers. The resulting file was split into two batches of fifty sentences, each of which was assessed by two experts. For each tagged word in the sample, the experts were asked to decide whether:

0: there was no mistake;

1: the lemma was correct, but not the part-of-speech;

2: the POS-tag was correct, but not the lemma;

3: both the POS-tag and the lemma were wrong;

4: there was a segmentation error (only for the CRF tagger).

At the end of the annotation process, the agreement between the two judges was computed for both batches. Since the tags are nominal and we have only two annotators, we applied the weighted kappa coefficient (Cohen, 1968) to measure agreement [2]. The results of the evaluation are reported in Section 4.

3.3. Computing lexical frequencies

Once the corpus tagged, the last step was to compute the word frequency counts per level and normalized them in various ways. The first normalization process that can be applied to the counts is simply to normalize the raw frequencies by level (*RFL*), since we do not have the same number of words per level. However, as noted by (Francis and Kucera, 1982), lower frequency words tend to be context specific, appearing in a small number of texts, but sometimes with a unusually high frequency within those texts. This finding has crucial implications when one wants to estimate counts from a textbook corpus (Lété et al., 2004). The content of textbooks is guided by a set of competencies and tasks related to various types of situations, which are defined only to a certain extent by the CEFR guidelines. Textbook designers therefore have quite a latitude to decide which topics will be included in their book. As a consequence, it is likely that the importance of some low frequency words, related to specific topics, will be overestimated using raw frequencies, especially when a topic generally encompasses several texts within the same lesson. To reduce this effect, we transformed the *RFL* using a dispersion index (D) as described in Carroll et al. (1971):

$$D_{w,K} = \log(\sum p_i) - [\sum p_i \log(p_i) / \sum p_i] / \log(I) \quad (1)$$

For a corpus with K levels of difficulty (in our case, $K = 6$), each of them including I textbooks or readers, the D of a given word w for the level K requires to use p_i, the probability that a word appears in the textbook i and I, which is the number of textbooks at the level k. When $p_i = 0$, $p_i \log(p_i)$ was also considered as 0. Once all Ds were computed, we finally combined the RFL with D to obtain U, the estimated frequency per million for a given word w. The formula is a follows (Carroll et al., 1971):

$$U = (1,000,000/N_k)[RFL * D + (1 - D) * f_{min}] \quad (2)$$

in which N_k is the total number of tokens for the level k and f_{min} represents $1/N$ times the sum of the products f_i and s_i, where f_i is the frequency of a word in the textbook i and s_i corresponds to the number of tokens in the textbook.

4. Results

Applying the above methodology, we obtained the first graded-lexicon for FFL that is compliant with the CEFR scale. In this section, we first investigate the quality of the tagging process (section 4.1.), then we describe the resource, which has been declined in two versions corresponding to the two taggers (section 4.2.). In the last part of this section, we further investigate the quality of the produced resource with some additional experiments.

4.1. Evaluation of the taggers

The first tagger assessed in this section is the TreeTagger. It is based on tree classifiers assisted by some lexical ressources and it has reached 96,36% accuracy on Penn-Treebank data for English (Schmid, 1994). Prior to its evaluation, we compared the expert agreement on both batches measured with the weighted kappa. As regards the interpretation of κ values, Artstein and Poesio (2008, 22) state: "CL researchers have attempted to achieve a value of κ (more seldom, of α) above the 0.8 threshold, or, failing that, the 0.67 level allowing for tentative conclusions". From this thumbrule, it appears that the expert agreement on both batches is good: 0.90 for the first batch and 0.83 for the second. For both batches, the two experts subsequently discussed about their divergences in order to settle a common annotation on which they agreed. This is the reference annotation we used to evaluate the quality of the tagging.

[2] The implementation used was from the NLTK python package (Bird et al., 2009).

As regards the second tagger, it is based on conditional random fields and large coverage linguistic resources. It has reached 97.34% F-measure on the French Treebank. The agreement scores for the evaluation process of this tagger are also very substantial: 0.84 for the first batch and 0.66 for the second. The fact that the κ scores are lower is partly due to the introduction of the fifth category : segmentation errors, since the detection of MWEs in this type of task is far from being obvious. As for the TreeTagger evaluation, all experts settle their divergences to define a reference annotation.

Once a reliable reference was obtained, we computed the proportion of the different types of errors for both taggers. Table 2 summarizes the results. First, it appears that the quality of the tagging is rather good in both cases (respectively with an accuracy of 94.2% and 95.8%), even though the corpus includes different types of texts, some of which are not usually used in the tagger's training corpus (e.g. dialogues or poems). This result is good news: the final resource presents an accuracy within those rates.

	TreeTagger	**K-ET Tagger**
correct	94.2%	95.8%
POS errors	2.6%	1%
Lemma errors	1.3%	0.5%
POS + lemma	1.9%	1.1%
Segmentation	/	1.6%

Table 2: Performance of the two taggers on our evaluation sample

When comparing the behavior of the two taggers, it is clear that TreeTagger makes more mistakes as regards part-of-speech categorization and lemma identification, some of which, however, are due to segmentation problems. For instance, TreeTagger may split a MWE, thus providing an erroneous analysis of its components. Furthermore, it is characterized by two features that proved problematic for our purposes. First, when no lemma is found, an *<unknown>* tag is produced, which means the loss of an occurrence for us. More importantly, TreeTagger sometimes outputs double lemmas, such as *être|suivre* ('to be' and 'to follow' for the French *suis*), when it cannot disambiguate between the two forms from the context. Double lemmas being obviously not a desirable feature for a lexicon, we had to take care of them with manual rules. Four cases of double lemmas were observed:

- double lemmas for verbs that actually have the same surface form (e.g. *étayer|étayer* 'support'). In this situation, we simply kept one of the lemma;
- singular and plural forms (e.g. *lunette|lunettes* 'telescope|glasses'). In this case, we selected the most common form (e.g. *lunettes*), since the competing form was generally quite rare and less relevant for a pedagogical resource;
- a masculine and a feminine form for the same word, usually a nominalized adjective (e.g. *anglais|anglaise* 'English'). In this case, we favoured the masculine form, except for some specific cases (e.g. *arrivé|arrivée* 'arrived|arrival');
- finally, some of the double lemmas were composed of two different verbs, some of the inflected forms of which are identical (e.g. *être|suivre* 'to be|to follow'). These cases were ignored, since counting an occurrence for both forms led to wrongly estimated frequency for pairs in which one of the form is very common whereas the other is quite rare (e.g.*être|sommer* 'to be|to summon') .

On its part, the CRF tagger makes less part-of-speech mistakes as well as wrong lemma identification than the TreeTagger. Its main shortcoming goes along with its main advantage, namely its ability to detect sequences of tokens likely to be MWEs. This creates segmentation problems for about $1,6\%$ of the tokens, in which case the tagger either misses an interesting MWE (e.g. *parti pris* 'prejudice' is split into two tokens), or more problematically, it creates a sequence of tokens that do not corresponds to a MWE (e.g. *parler d* 'speak ab'). Even though this second kind of errors is rare (less than 1%), due the size of the corpora, it occurs enough to create several hundred of erroneous entries that rendered necessary a manual verification of the resource.

4.2. One resource, two versions

The comparison and evaluation of our two taggers showed that the version of the resource produced with TreeTagger (FLELex_TT) was cleaner as regards the entries, but likely to estimate frequency distributions slightly less correctly. On the other hand, the resource based on the CRF tagger (FLELex_CRF) provided better frequency estimations (due to the enhanced tagging process) and presented more entries (namely compound words and MWEs), but some of them were wrongly tokenized. Taking into account all these considerations, we nevertheless decided to distribute the two versions of the resource, giving complete user choice. Both lexicons can thus be used as pedagogical resources for teaching purposes. For iCALL and NLP tasks (text simplification or readability assessment), FLELex_TT might be better suited, provided that other tools compatible with TreeTagger tags are used. Since FLELex_CRF was manually cleaned and provides a richer list of entries, it should be considered as the reference version of FLELex.

The TreeTagger-based version of FLELex includes 14,236 entries, while the CRF-based version includes as much as 17,871 entries [3]. This difference is obviously due to the ability of the CRF-tagger to detect MWEs. All entries in the lexicon are presented along with their POS tag, a U frequency for each of the six levels of the CEFR and the U frequency computed on the whole corpus.

Table 3 illustrates the type of information contained in FLELex, presenting the entries for *voiture* (1) 'car', *abandonner* (2) 'forsake, give up', *justice* (3), *kilo* (4) and *logique* (5) 'logic'. We can see that concrete concepts (such

[3]Both resources will be made available to the community at the following address: `http://cental.uclouvain.be/flelex`

as *kilo* and *voiture*) are mainly related to the first stages of the learning process (A1 to B1) and then tend to be less used in later stages. On the contrary, *justice* and *logique* are terms typical of more advanced levels, while *abandonner* has a more uniform distribution. As for MWEs, *en bas* 'at the bottom' appears to be a more common expression than *en clair* 'clearly'. Similarly, the prepositional group *sous réserve de* 'subject to' appears in a mid-level but is not used elsewhere. Needless to say that this type of information could prove useful in various pedagogical contexts, especially in iCALL applications.

4.3. Further analysis of the resources

This section reports results on further investigations about FLELex data. The distribution of the various part-of-speech categories in both versions is first detailed on table 4:

POS	TTagger		CRF Tagger	
NOUNs	7,837	55.05%	9,083	50.83%
ADJs	3,015	21.18%	3,453	19.32%
VERBs	2,598	18.25%	2,763	15.46%
ADVs	603	4.24%	1,534	8.52%
Total	14,053	98.72%	16.833	94.13%
Other categories	183	1,29%	1,038	5,81%
Total	**14,236**		**17,871**	

Table 4: Distribution of lexical units by POS, in the TreeTagger and CRF versions of FLELex.

It is interesting to note that the proportion of adverbs doubles in the CRF-based version. Since adverbs are a category limited in size, this finding must be interpreted as the fact that about 900 adverbial phrases were detected by the CRF tagger. We assume that this type of expressions are very valuable for language learning purposes. As regards the verb category, the figures in both cases are pretty similar, which probably means that verbal MWEs were not much detected by the CRF tagger. This seems logical, since verbals MWEs are prone to be separated by a complement and are therefore much more challenging to detect.

Another interesting insight is the number of words that were seen only once in the corpus (hapax). In the TreeTagger-based version, 33% of the entries are hapaxes in terms of raw frequencies and only 26% of the entries have a frequency higher than 9, while words having a raw frequency higher than 100 amounts to 4%. In the CRF-based version, only 20% of the entries are hapaxes, 31% of the entries have a frequency of 10 or higher, while 6% of the entries exceeds 100 occurrences. Since there are more entries in the second version, it is surprising to obtain such figures, the opposite behavior being expected. It is likely that the phenomenon of double lemmas in TreeTagger is partly responsible of a loss of occurrences, but it does not explain everything. Another issue raised by these figures is that the corpus might be too small to provide a robust estimation of the frequencies by level for the less frequent words in the database.

To investigate these issues, we performed a final test on the TreeTagger-based FLELex. All entries were compared with those of a general French lexicon providing frequencies, namely Lexique3 (New et al., 2001). This resource includes 47,342 lemmas along with a large set of psycholinguistic features. One of these features is the lemma frequency, estimated on a corpus of film subtitles amounting to about 50 million words (New et al., 2007). With this lexicon, it was possible to (1) check whether the entries from FLELex indeed correspond to existing entries and (2) to compare the frequencies estimated on our small corpus with those computed on a much larger dataset in order to assess the robustness of our frequency estimation process.

Interestingly, we found 622 entries in FLELex that are not listed in Lexique 3. Some of them are real words missing from Lexique 3 (e.g. *marquise* (noun) 'marquise' or *oxydant* (adjective) 'oxydizing', while others correspond to a tagging error that has produced an incorrect combination of lemma and POS tag (e.g. *barbe* (adjective) 'beard'). Manually investigation of these cases appears as an interesting perspective. However, this analysis also shows that tagging errors have yielded only a limited number of wrong entries in FLELex.

As regards the frequency estimation issue, we compared the U values in FLELex with the frequencies in Lexique 3 using a Pearson correlation. The correlation reaches 0.84, which proves that our frequencies are comparable to those of Lexique 3, estimated on a much larger corpus. Furthermore, the differences observed between the two resources do not necessarily have to be attributed to the smaller size of the corpus, since it is expected that the distribution of words in textbooks does no follow exactly the distribution of words in a corpus of film subtitles.

5. Conclusion

In this paper, we have presented the first graded lexicon for FFL that reports frequencies by level, according to the CEFR scale. The resource has been built from a corpus of 777,000 words from available textbooks intended for FFL learners and distributed among the six CERF levels. The electronic version of the corpora has been tagged using two different tools enabling to obtain two graded-lexicons. The first tagger presents an overall accuracy of 94.2%, whereas the second has an overall accuracy of 95.8%. Moreover, this CRF tagger is able to identify MWEs, although it sometimes fails, tokenizing wrong MWEs (wrong entries have been manually removed afterwards).

The different tagging strategies entail a difference in the resulting data (in terms of size and nature). We thus propose two versions of the same resource that will be freely provided to the community to be used for different purposes: for humans (as lexicons in assistive tools for writing, in educational activities for learning vocabulary) and in NLP tasks (automatic assessing the lexical difficulty of a FFL text, selecting simpler words within text simplification systems, etc.).

In future work, we plan to enhance the coverage of the resource and the lexical information associated to the entries. We will also compare the two versions in different NLP applications addressed to different users.

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

Table 3: Example of some entries in the CRF-based FLELex: (1) 'car', (2) 'forsake, give up', (3) 'justice', (4) 'kilo', (5) 'logic', (6) 'at the bottom', (7) 'clearly' and (8) 'subject to'.

Acknowledgements

This project is partly financed by the Programme Hubert Curien (PHC) Tournesol 2014 (France-Fédération Wallonie-Bruxelles).

6. References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bahns, J. and Eldaw, M. (1993). Should We Teach EFL Students Collocations? *System*, 21(1):101–14.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., and Loftis, B. (2007). The english lexicon project. *Behavior Research Methods*, 39:445–459.

Beacco, J.-C. and Porquier, R. (2007). *Niveau A1 pour le français (utilisateur/apprenant élémentaire)*. Didier, Paris.

Beacco, J.-C., Lepage, S., Porquier, R., and Riba, P. (2007). *Niveau A2 pour le français (utilisateur/apprenant élémentaire) niveau intermédiaire*. Didier, Paris.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

BNC-Consortium. (2001). The british national corpus, version 2 (bnc world).

Brysbaert, M., Lange, M., and Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1):65–85.

Buchanan, M. A. (1927). *A graded Spanish word book*. The University of Toronto Press.

Carroll, J., Davies, P., and Richman, B. (1971). *The American Heritage word frequency book*. American Heritage Publishing Co., New York.

Catach, N. (1984). *Les listes orthographiques de base du franais*. Nathan. Collection Nathan Recherche, Paris.

Cattell, J. (1885). The inertia of the eye and brain. *Brain*, 8:295–312.

Church, K. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16:22–29.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In *In Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213–220.

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Hatier, Paris.

Constant, M. and Sigogne, A. (2011). Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56. Association for Computational Linguistics.

Content, A., Mousty, P., and Radeaux, M. (1990). Brulex : une base de données lexicale informatisée pour le français écrit et parlé. *L'année Psychologique*, 90:551–566.

Deville, G., Dumortier, L., and Meurisse, J. R. (2013). Ressources lexicales pour l'aide à l'apprentissage des langues. In *Ressources lexicales: contenu, construction, utilisation, évaluation*, volume 30, pages 291 – 311. John Benjamins, Amsterdam, Gala, N. et Zock, M. edition.

Dias, G., Guilloré, S., and Lopes, J. (2000). Extraction automatique dassociations textuelles à partir de corpora non traités. In *Proceedings of 5th International Conference on the Statistical Analysis of Textual Data*, pages 213–221.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Mot, A., Augustinova, M., and Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496.

Francis, W. N. and Kucera, H. (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, New York.

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 *E-lexicography in the 21st century: thinking outside the paper.*, Tallin, Estonia.

Gougenheim, G. (1958). *Dictionnaire fondamental de la langue française*. Didier, Paris.

Henmon, V. A. C. (1924). *A French word book based on a count of 400,000 running words*. Bureau of Educational Research, Madison: University of Wisconsin.

Howes, D. and Solomon, R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

Kucera, H. and Francis, W. N. (1967). Computational analysis of present-day american english.

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

Lively, B. and Pressey, S. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9:389–398.

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

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In Besner, D. and Humphreys, G., editors, *Basic processes in reading: Visual word recognition*, pages 148–197. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

New, G. A., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur Internet : Lexique 3. *L'année psychologique*, 101:447–462.

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.

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

Selva, T., Verlinde, S., and Binon, J. (2004). Alfalex, un environnement daide à lapprentissage lexical du français langue étrangère. In *Congrès de l'ACFAS*, Montréal.

Solomon, R. and Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3):195–201.

Thorndike, E. and Lorge, I. (1944). *The Teacher's Word Book of 30,000 words*. Teachers College, Columbia University, New York.

Thorndike, E. (1921). *The Teacher's Word Book*. Teachers College, Columbia University, New York.

Vander Beke, G. E. (1932). *French word book*, volume 15. Macmillan.

Vogel, M. and Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.