EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language

Dürlich, L.[1,2]**, François, T.**[2,3]
[1] FAU Erlangen-Nürnberg,
[2] CENTAL/IL&C, University of Louvain
[3] Chargé de recherche FNRS
luise.duerlich@fau.de, thomas.francois@uclouvain.be

Abstract

This paper introduces EFLLex, an innovative lexical resource that describes the use of 15,280 English words in pedagogical materials across the proficiency levels of the European Framework of Reference for Languages. The methodology adopted to produce the resource implies the selection of an efficient part-of-speech tagger, the use of a robust estimator for frequency computation and some manual post-editing work. The content of the resource is described and compared to other vocabulary lists (MRC and BNC) and to a reference pedagogical resource: the English Vocabulary Profile.

Keywords: CEFR-graded lexicon, English as a foreign language, vocabulary

1. Introduction

The strong relation between vocabulary knowledge and reading comprehension has been thoroughly researched and confirmed (Laufer, 1992). Second language acquisition (SLA) research has established that readers should know between 95% and 98% of the words in a text to adequately comprehend it (Laufer and Ravenhorst-Kalovski, 2010). Given that the larger the vocabulary, the better the comprehension of texts, foreign language curriculums should stress the need to teach new vocabulary items. A larger vocabulary indeed means being able to understand a larger scope of texts. However, analyses of large corpora indicate that, from kindergarten through college, native speakers encounter approximately 150,000 different words (Zeno et al., 1995). For learners, it is impossible to pick up all of them. Designers of foreign language curriculums, publishers of educational materials and textbooks, or even teachers are thus 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 a learner audience and (2) to rely on expert knowledge, such as teacher expertise or linguists' recommendations. Following the first approach, frequency lists have been derived from various well-known corpora of English. The first significant vocabulary list for English included 20,000 words and was laboriously produced by Thorndike (1921) without the help of any computer. It was later extended to 30,000 words by Thorndike and Lorge (1944). The first computational list was obtained by Kučera and Francis (1967) from the Brown corpus and has a large influence in education and psychology. More recently, other lists have been developed from larger corpora, such as the CELEX database (Baayen et al., 1993), the Zeno list (Zeno et al., 1995), the list based on the British National Corpus (BNC) (Leech et al., 2001), or SUBLEX (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.

The second approach to set a learning path for vocabulary has relied on expert knowledge. The most popular resources of this kind are the Reference Level Descriptions (RLDs). They are based on the scale of the Common European Framework of Reference for languages (CEFR), published by the Council of Europe (2001), and ranging from A1 to C2. The CEFR describes the skills learners should develop at each of the six proficiency level of the scale. For instance, for reading skills, a A1 learner should understand very simple sentences and familiar words and names, while a B1 learner should understand texts consisting mainly of high frequency language as well as the description of events, feelings, wishes in personal letters. At C2, a learner is able to read virtually all forms of written language however abstract, structurally or linguistically complex. However, such descriptions remain elusive and the limitations of the CEFR for practical purposes have been stressed (North, 2005, 40).

RLDs aims at providing more detailed linguistic guidelines based on the CEFR descriptors for over 20 European languages. They consist in lists of words, multi-words, functions, or syntactic structures where each entry is connected to one level of the CEFR. Following the *Guide for the production of RLD* published by the Language Policy Division DG IV of the Council of Europe in Strasbourg, the RLDs have been elaborated on the basis of inventories of statistical frequencies, perusal of large bodies of texts, expert knowledge, and learners' productions (Marello, 2012, 328). The RLD for English has been developed within the English Vocabulary Profile project, or EVP (Capel, 2010; Capel, 2012), using a corpus-informed approach based on learner production from the *Cambridge Learner Corpus*. The methodology applied has the great advantage of being able to assign different difficulty levels to the different senses of a word. However, Alderson (2007) stressed that relying almost entirely on the *Cambridge Learner Corpus*,

Genres	A1		A2		B1		B2		C1		Total	
Ad	14	(544)	11	(943)	26	(2,339)	7	(1,027)	8	(969)	66	(5,822)
Dialogue	52	(2,643)	51	(3,671)	77	(7,438)	45	(6,904)	27	(3,997)	252	(24,653)
E-Mail	1	(15)	14	(859)	32	(2,886)	18	(2,695)	7	(478)	72	(6,933)
Informative	27	(1,921)	47	(5,711)	123	(22,271)	104	(24,821)	110	(26,542)	411	(81,266)
Mail	1	(68)	4	(180)	12	(1,175)	6	(765)	2	(300)	25	(2,488)
Narrative	51	(4,285)	62	(7,837)	24	(4,487)	27	(6,751)	11	(1,835)	175	(25,195)
Reader Text	51	(21,476)	139	(69,610)	32	(28,494)	192	(88,651)	135	(92,715)	549	(300,946)
Recipe	-	-	-	-	3	(214)	-	-	1	(84)	4	(298)
Sentences	59	(2,310)	37	(2,185)	47	(3,246)	17	(1,571)	20	(1,921)	180	(11,233)
Various	16	(927)	39	(3,245)	102	(10,340)	43	(7,401)	37	(5,431)	237	(27,344)
Total	275	(34,422)	407	(94,460)	478	(82,890)	459	(140,586)	358	(134,272)	1,971	(486,178)

Table 1: Text and word distributions throughout the corpus by level and type of texts.

which is a collection of performances on Cambridge examinations, may be an issue for generalization. Moreover, RLDs are not able to discriminate from all the words related to a given proficiency level, the most important ones to learn.

Since 2014, a third approach has been investigated within the CEFRLex project, which adopts an original view on lexicon learning. In contrast with the classic approach that models lexical knowledge in a nominal fashion (i.e. each word is assigned to a given level of proficiency, which implies that all learners from a given level should know all words from this level), the CEFRLex project assumes a continuous vision of lexical learning. Words are rather described in terms of a frequency distribution over the CEFR proficiency levels. This distribution is obtained from collections of pedagogical documents that are intended for L2 learners and labelled in accordance with the CEFR scale. We are therefore able to describe the usage of words in a more subtle way, for instance stressing the fact that a given word, usually considered to be learned at a specific level (e.g. B1), may already occur in texts intended for lower proficiency levels. Similarly, the availability of frequencies per level allows to rank words assigned to the same level of proficiency by frequency.

The CEFRLex project aims to bring together various research teams across Europe to produce vocabulary lists for several European languages. So far, lists for French (François et al., 2014) and Swedish (François et al., 2016) have been developed, and the current paper introduces the English version of the CEFRLex list, called EFLLex. The next section describes the methodology applied to create the English list and to estimate the frequencies from the labelled corpus. Section 3. then describes the content of the resource and compares it with similar resources.

2. Methodology

The development of EFLLex draws on the methodology used for the other lists of the CEFRLex project, but it also met some specific challenges. The general principle is the following: based on a labelled corpus, where each text has been assigned a CEFR level by human annotators (e.g. by the authors of the textbooks or of the simplified readers), the frequency of each word is estimated and normalized for each CEFR level, which results in the word distribution over the CEFR levels. In this section, we first describe the collection process of the labelled corpus; we then discuss the choice of the automatic tagger used, before describing the frequency estimation technique. The section concludes with some considerations about the manual correction of the resource.

2.1. Source corpus

We collected a corpus consisting of 1,971 texts extracted from 17 textbooks, 33 graded readers and 7 online materials designed for learners of English as a foreign language. These books come from popular publishers such as Cambridge University Press, Oxford University Press, and Exam English Ltd. All texts selected for the study were related to a reading comprehension task and were classified according to the CEFR levels by the publishers' teams. Very few of the available resources were actually assigned the level C2; our corpus thus only includes texts within the levels A1 to C1. As a whole, the corpus contains 486,178 words that are distributed across the levels as displayed in Table 1. This table also describes the distribution of texts over different types of texts (e.g. dialogue, informative, narrative, mail, etc.), highlighting the variety of texts that have been collected. It should be mentioned that more advanced levels (B2 and C1) account for the biggest subcorpora, as texts of these levels are more easily available and generally longer.

2.2. Tagging

As a next step, the corpus had to be lemmatized and part-of-speech (POS) tagged, for the two following reasons. First, in the CEFRLex project, the choice has been made to compute frequencies of lemmas observed in the corpus instead of frequencies of inflected forms. Counting inflected forms 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 in texts 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. The rationale to use POS tags is that it enables to disambiguate homographic forms with different part-of-speech tags.

The choice of the POS-tagger is also critical, as lemmatization or tagging errors produce unwanted effects on the data such as:

- entries with wrong part-of-speech tag (e.g. *much* is tagged as JJ (adjective) instead of RB (adverb) or *fatigue* is tagged as ":" instead of NN for noun);
- entries with an incorrect or at least questionable lemma (e.g. *she* instead of *her*, *1* instead of *first*);
- tags that can be correct, but are erroneous in the specific context of the word (e.g. to tag the word *quiet* in *the peace and quiet* as an adjective)[1].

We also wanted to use a system able to detect phrasal verbs and multi-word expressions, as they are of prime importance in the L2 learning process (Paquot and Granger, 2012).
Based on the previous constraints, the following five POS-taggers were compared to find the most suitable one for the task:

- the TreeTagger (Schmid, 1994),
- a HMM-tagger based on the TnT-tagger (Brants, 2000) provided by the FreeLing library for C++ (Carreras et al., 2004),
- the left-to-right model of the Stanford Log-linear POS Tagger (Toutanova et al., 2003),
- the tagger provided by the SVMTool (Giménez and Màrquez, 2004)
- and a tagger based on automatic feature extraction provided by the NLP4J module (Choi, 2016).

To compare the taggers, we assessed their performance on one hundred sentences randomly sampled from the corpus. Two human experts manually checked the annotation output of each tagger and assigned each tagged word one of the five following categories: **(0)**: correct lemma and POS-tag, **(1)**: correct lemma, but wrong POS-tag, **(2)**: wrong lemma, but correct POS-tag, **(3)**: wrong lemma and wrong POS-tag, **(4)**: segmentation error (e.g. *it-* or *'pages*), which usually leads to tagging or lemmatization errors. The agreement between the two human experts was calculated using Cohen's Kappa. Agreement scores varied from 0.627 for the SVMTool tagger to 0.378 for the Freeling tagger. Such levels of agreement are weak, so an additional step was included, in which both experts compared their annotations in order to agree on a reference annotation, on which the performance scores presented in table 2 were calculated.

Category	TreeTagger	HMM	Stanford	SVMTool	NLP4J
(0) Correct	92.74%	95.35%	95.34%	93.49%	**95.88%**
(1) POS	5.72%	3.15%	4.45%	4.67%	3.86%
(2) Lemma	0.14%	0.29%	0.07%	0.78%	0.06%
(3) Lemma + POS	0.84%	0.72%	0.14%	0.28%	0.0%
(4) Segmentation	0.56%	0.5%	0.0%	0.78%	0.19%

Table 2: Evaluation of the POS-Tagger results

[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.

We observed that none of the taggers correctly identified indefinite and demonstrative pronouns, nor some adverbs of time that could also be tagged as nouns, such as *today* or *tomorrow*. Furthermore, there were several issues when it comes to lemmatization, e.g. *me* was lemmatized as *I* by the Standford Tagger and the lemma *she* was assigned to *her*. Since the NLP4J Tagger tagged most of the words correctly, we decided to use this tagger on the whole corpus. Using the entire parsing output of NLP4J, that featured dependency labels and verb particles, it was also possible to reconstruct compounds and phrasal verbs in the corpus using specific rules.

2.3. Estimating lexical frequencies

Similarly to what was done for the other resources of the CEFRlex family, the normalized frequency per million words is obtained by first computing the raw frequencies by level (*RFL*) based on the corpus. In a second step, the *RFL* are weighted by a dispersion index (D), intended to counteract the effect of context-specific low frequency words being overused within a small number of texts. As noted by Francis and Kučera (1982), lower frequency words tend to be context-specific: they appear in a small number of texts, but sometimes with a unusually high frequency within those texts. This observation is of particular relevance when estimating counts from textbooks. Their content is guided by a set of competences 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 adapted the *RFL* using a dispersion index (D), adapted from Carroll et al. (1971):

$$D_{w,k} = \frac{\log\left(\sum p_i\right) - \frac{\sum p_i \log(p_i)}{\sum p_i}}{\log(I)} \tag{1}$$

In a corpus of K difficulty levels, each containing I text sources (e.g. textbooks), in order to obtain the dispersion index D of a word w for a level k, we rely on two pieces of information: (1) p_i the probability of the word w in the i^{th} source, which is computed as the frequency of w in the source i divided by s_i, the number of words in the source i (if $p_i = 0$, $p_i log(p_i)$ was considered as equal to 0); (2) I is the total number of sources.
Combining the dispersion index D and *RFL* as follows, we obtain U, the frequencies per million words:

$$U_{w,k} = \frac{1,000,000}{N_k}(RFL * D_{w,k} + (1 - D_{w,k}) * f_{min}) \tag{2}$$

where N_k is the number of words in level k and f_{min} is defined as:

$$f_{min} = \frac{1}{N_k}\sum_{i=1}^{I} s_i f_i \tag{3}$$

lemma	tag	A1	A2	B1	B2	C1	total
cat	NN	2.940	202.796	31.681	33.339	28.9847	65.019
empty	JJ	86.492	150.888	65.947	194.801	123.405	156.021
explore	VB	20.578	54.677	73.625	46.070	56.961	69.590
obviously	RB	0	11.034	2.589	68.463	36.665	30.689
tiresome	JJ	0	0	0	0.315	0.815	0.611
video	NN	2.467	0.556	34.825	23.802	13.248	18.431
write	VB	934.708	378.337	760.734	536.380	713.326	549.909
shopping centre	NN	0	5.040	2.589	0	0.815	1.946
sign up	VB	0	0.887	10.789	2.499	6.216	5.302

Table 3: Example of some entries in EFLLex.

Level	# entries	# new entries	Hapax	>10	EVP	FLELex	SVALex
A1	2,395	2,395	893	509	601	4,976	1,157
A2	4,205	2,478	1,633	1,000	925	3,516	2,432
B1	5,607	2,740	2,366	1,003	1,429	4,970	4,332
B2	8,228	3,935	3,580	1,571	1,711	1,653	4,553
C1	9,232	3,733	4,254	1,591	N/A	2,122	3,160

Table 4: Distribution of EFLLex entries per CEFR level, including the number of attested entries per level, number of new entries, number of hapax legomena, number of words occurring more than ten times. We also provide the number of new entries for English Vocabulary Profile (EVP) (Capel, 2010), the number of new items in FLELex (François et al., 2014) and in SVALex (François et al., 2016) for comparison.

with f_i, the frequency of a given word in a source i, and s_i, the number of word in i, for each source divided by N_k.

2.4. Manual correction

The automatic creation of the EFLLex list was followed by a phase of manual correction, where we compared our resource with the MRC machine usable dictionary (Wilson, 1988) to find any combination of lemma and POS unattested by the MRC dictionary. The resulting list of ”new” pairs was then checked by hand in order to decide whether we were dealing with some kind of tagging or lemmatization error or with correct pairs that, for some reason, were not included in the MRC dictionary. Among the entries absent in the MRC list, we found some recent expressions - e.g. words reflecting new technologies and media such as the verb *to blog* (as well as the nouns *blog* and *blogger*), *internet* or *smartphone* - and - as was quite often the case - compound words (e.g. *food poisoning*) and phrasal verbs (e.g. the verb *to get up*) that the MRC list did not record. The majority of the ca. 6,600 words had to be checked manually. Most of these entries were correct words and only 1,104 were ”real” errors that could either be traced back to errors in the raw corpus or to errors made by the tagger.

3. Description of the resource

EFLLex contains 15,280 entries. Examples of entries can be seen in Table 3. Each entry corresponds to the combination of a lemma and a part-of-speech: the frequency of the lemma at each level of the CEFR (C2 excepted) is described, as well as its total frequency in the corpus. From Table 3, it can be seen that EFLLex includes both multiword expressions such as *shopping center* or phrasal verbs such as *sign up*. As already mentioned above, multi-word expressions are of utmost importance for L2 teaching purposes (Bahns and Eldaw, 1993), but are absent from most of the English frequency lists that were introduced in Section 1.

The distribution of the entries across the levels A1 to C1 is displayed in Table 4. As can be observed, the total number of words (*# entries*) encountered at a given level increases from elementary to more advanced levels, as could have been expected. If we take into account the number of new entries per level (*# new entries*), there is also a moderate increase as the levels advance. This was expected, but if we compare these figures with those from FLELex and SVALex (columns 7 and 8), an interesting pattern emerges. Compared to SVALex, words are encountered faster, i.e. more words are introduced at the lower levels (A1 and A2), but this trend reverses at the intermediate levels (B1 and B2). In contrast, in FLELex, words are introduced faster at the elementary levels. As already mentioned in (François et al., 2016) we believe that this pattern results from the variation in corpus size. The results obtained for EFLLex confirm that the larger the corpus used to estimate the frequencies, the quicker new entries are introduced.

Besides comparisons with other resources from the CEFRLex project, it is also interesting to compare EFLLex with the EVP referential. As can be seen from Table 3, the EVP introduces far fewer words per level, which could be explained by two main reasons. First, the EVP aims to cover learners’ productive vocabulary, which is known to be smaller than the receptive vocabulary covered by EFLLex. Second, EFLLex includes both words from the core vocabulary that should be taught at a given level, and peripheral words that can be encountered by learners, but should probably be learned at a later stage of the curricu-

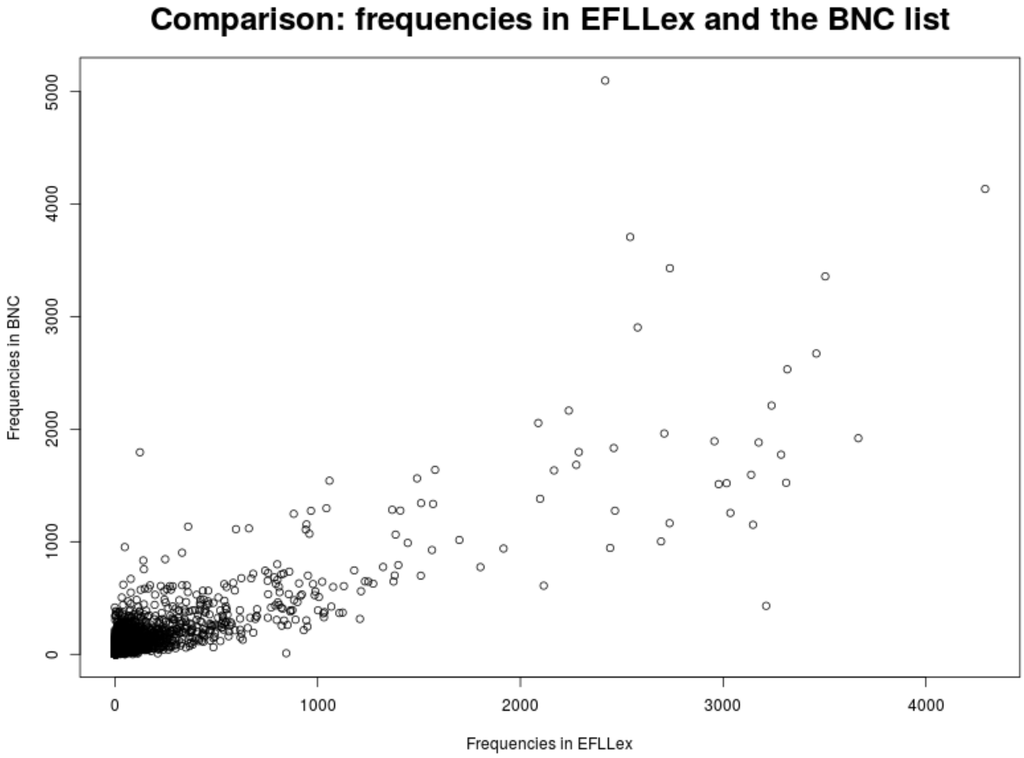

Figure 1: Association between the frequencies in the BNC list and EFLLex.

lum. In other words, EFLLex favours coverage over precision. EFLLex is therefore a useful tool to get a good idea of a word usage within EFL pedagogical materials, although further investigations are needed to interpret the frequencies in terms of teaching goals.

As regards frequency estimation, the limited size of the corpus has implications on the frequency distributions of entries, in a similar way to the other resources of the CEFR-Lex project. There is quite a significant number of words occurring at one level only, as revealed by the number of hapax legomena in the resource (column 4). Moreover, only 3,051 entries have a total frequency higher than 10 occurrences per million words, which is a clear limitation of the resource. Therefore, conclusions on the real usage of infrequent entries in EFLLex should be drawn with caution.

To better investigate this issue, the frequencies obtained on all levels (**total**) were compared with two popular frequency lists for English: (1) the BNC list and (2) the MRC machine usable dictionary. The BNC list consists of 678,211 entries for lemmas reporting frequencies per million words based on both written and spoken parts of the BNC (Leech et al., 2001). As the qualities of the BNC as a corpus are well acknowledged, it is reassuring to observe that the Pearson product-moment correlation coefficient between frequencies of EFLLex and the BNC list is very high ($r = 0.97; p < 0.0001$). Figure 1 shows this association for words having a frequency lower than 5,000 in EFLLex. For the counts from the London-Lund Corpus of English Conversation taken from the MRC list, the correlation is notably weaker ($r = 0.53; p < 0.0001$). Comparison with frequencies reported by the MRC list, however, is problematic in that frequencies have been added together over all possible POS-tags a word form can take. This means that the word *laugh*, for example, is assigned the same frequency, whether it be a verb or a noun. Thus, the MRC does not report reliable frequency information for specific token and tag combinations. Table 5 illustrates, with the word *well*, the consequences of such a choice when carrying out a correlation analysis such as ours. In conclusion of this experiment, the high correlation with the frequencies in the BNC list demonstrates that, even though the size of our corpus was limited, frequencies in EFLLex seem to offer a good estimate of word usage.

POS	EFLLex	BCN	MRC
well JJ (adjective)	5.9	42	1,753
well NN (noun)	1.7	14	1,753
well RB (adverb)	661.0	1,119	1,753
well UH (interjection)	787.4	-	1,753
well VB (verb)	-	2	1,753

Table 5: Frequencies of *well*, depending on its part-of-speech, in EFLLex, the BNC and the MRC lists.

4. Conclusion

In this paper, we have introduced EFLLex, a very original lexical resource for English that describes word usage in pedagogical materials over the CEFR levels. We have reported the methodological details of the development of

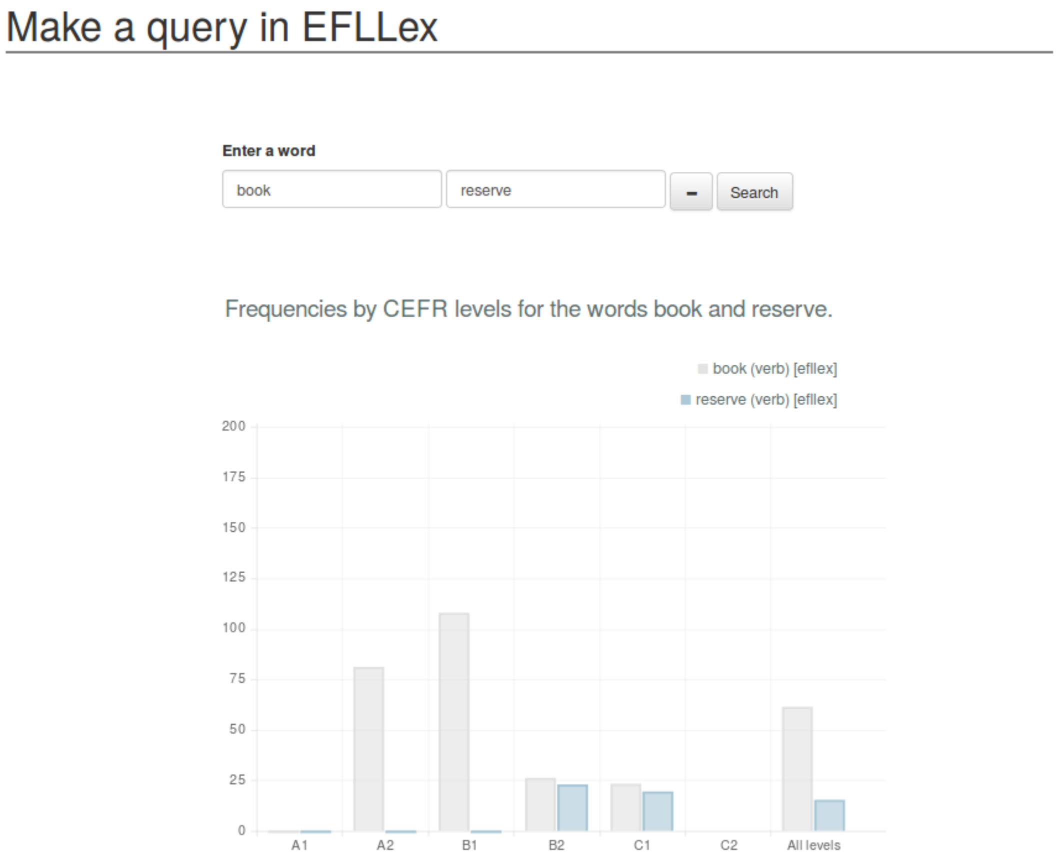

Figure 2: Screen capture of EFLLex website, showing the distributions of the verbs *book* and *reserve*.

this resource, which is part of a larger project, the CEFR-Lex project that aims to cover a wide range of European languages. The list is freely available for the scientific community[2] and can be queried through a web interface that displays the distribution of a word. The interface can also be used to compare the distributions of two words, as exemplified by the screen capture of this interface (see Figure 2). Finally, we plan to develop an additional tab for this interface that will allow to assess the vocabulary of a whole text: each word of the text that is included in EFLLex will be assigned one of the CEFR level based on its EFLLex distribution. Such tool draws from a similar system developed for French (Tack et al., 2016) and is an interesting alternative to the Text Inspector system[3] based on EVP.

In contrast with classic frequency lists, EFLLex is based on pedagogical texts intended for EFL learners and therefore represents a better image of the lexicon they encounter during their curriculum. It also offers a finer view of word use within a level: for instance, EFLLex makes obvious that *write* is a much more prevalent word at A1 (934 occurrences) than *explore* (20 occurrences). However, the resource also has some limitations as regards frequency estimation. Not only, the size of the corpus was limited, as there is a limited amount of pedagogical materials available, but it is also unbalanced in favour of more advanced levels. Still, as was shown by a comparison with the BNC list, the EFLLex frequencies are a robust estimation of the use of English words by learners. We believe EFLLex can provide interesting insights for language learning and will hopefully inspire more contributions to the domains of computer-assisted language learning and language teaching.

5. Acknowledgements

We would like to express our gratitude to Baptiste Degryse and Dorian Ricci for their invaluable help with the EFLLex interface. We also need to acknowledge the support of Alice Panier that helped us to collect more texts !

6. Bibliographical References

Alderson, J. (2007). The CEFR and the need for more research. *The Modern Language Journal*, 91(4):659–663.

Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). The CELEX lexical data base on CD-ROM.

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

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231.

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 im-

[2]http://cental.uclouvain.be/cefrlex/efllex/

[3]https://textinspector.com/

proved 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.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3:1–14.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

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

Choi, J. D. (2016). Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of NAACL*.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014)*.

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 the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219.

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

Giménez, J. and Màrquez, L. (2004). SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Laufer, B. and Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22(1):15–30.

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.

Marello, C. (2012). Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages). In *Proceedings of the XV Euralex International Congress*, pages 328–335.

North, B. (2005). The CEFR levels and descriptor scales. In *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference*, pages 21–66.

Paquot, M. and Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32:130–149.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 44–49.

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 the Tenth conference on International Language Resources and Evaluation (LREC'16)*, pages 230–236.

Thorndike, E. and Lorge, I. (1944). The teacher's word-book of 30,000 words.

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

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine Usable Dictionary, Version 2.00. In *Behaviour Research Methods, Instruments & Computers*, volume 20, pages 6–10. Springer US.

Zeno, S., Ivens, S., Millard, R., and Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates, Inc., New York.