Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 346–355
Marseille, 11–16 May 2020

Using Multilingual Resources to Evaluate CEFRLex for Learner Applications

Johannes Graën[1,2], David Alfter[1], Gerold Schneider[3,4]
[1]Språkbanken Text, Department of Swedish, University of Gothenburg
[2]Grael, Department of Translation and Language Sciences, Pompeu Fabra University
[3]Department of Computational Linguistics, University of Zurich
[4]English Department, University of Zurich
{johannes.graen, david.alfter}@svenska.gu.se, gschneid@es.uzh.ch

Abstract

The Common European Framework of Reference for Languages (CEFR) defines six levels of learner proficiency, and links them to particular communicative abilities. The CEFRLex project aims at compiling lexical resources that link single words and multi-word expressions to particular CEFR levels. The resources are thought to reflect second language learner needs as they are compiled from CEFR-graded textbooks and other learner-directed texts. In this work, we investigate the applicability of CEFRLex resources for building language learning applications. Our main concerns were that vocabulary in language learning materials might be sparse, i.e. that not all vocabulary items that belong to a particular level would also occur in materials for that level, and, on the other hand, that vocabulary items might be used on lower-level materials if required by the topic (e.g. with a simpler paraphrasing or translation). Our results indicate that the English CEFRLex resource is in accordance with external resources that we jointly employ as gold standard. Together with other values obtained from monolingual and parallel corpora, we can indicate which entries need to be adjusted to obtain values that are even more in line with this gold standard. We expect that this finding also holds for the other languages.

Keywords: Computer-Assisted Language Learning (CALL); Lexicon, Lexical Database; Multilinguality

1. Introduction

Graded vocabulary lists have different applications such as serving as a basis for textbook writers, learner dictionaries or as self-paced learning tool for language learners (Kilgarriff et al., 2014). Especially in a second language learning context, vocabulary knowledge is highly correlated with general language proficiency and is a prerequisite for successful communication (Nation, 2013).

The main problem is that most graded vocabulary lists do not contain an evaluation of their quality and reliability. Nevertheless, as an user of such resources, one might want to know how reliable the resource is before employing it in the context of a language learning application.

The Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2001) is a scale of proficiency divided into three broad levels, A, B, and C, each of which is further subdivided into two sub-levels, so that the full scale ranges over 6 levels, from A1 for beginning learners over A2, B1, B2, C1 to C2 for near-native learners.

The most prominent use of the CEFR is in the form of (1) language certificates such as the Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS), and (2) CEFR-graded textbooks. While most tests have their own scoring system, they can all be mapped onto the CEFR scale. CEFR levels are also used in classroom language teaching to differentiate between different learner groups. Thus, one can have a Swedish class for B1 learners, which presupposes that learners taking the class have mastered all or most of the skills of the lower levels A1 and A2 and should have mastered all or most skills introduced at B1 after having finished the class.

In this paper, we explore multiple hypotheses relating to the CEFRLex family, a collection of similar resources derived from CEFR-graded textbook corpora, which is available for several languages. Our first hypothesis is that similar words in two languages, i.e. good direct translations, should have similar CEFR levels. However, this also raises the question of culture- and language-specific vocabulary. A second hypothesis is that the broader a concept is, the lower its CEFR level should be, as the possibility of knowing at least one of the possible interpretations is higher than with highly specific vocabulary. Thirdly, we also explore how the frequency as reflected in CEFRLex and textbooks relates to the frequency of expressions in actual language by looking at the British National Corpus (The BNC Consortium, 2007) and the International Corpus of Learner English (ICLE) (Granger et al., 2009).

The CEFRLex resources are based on CEFR-graded textbooks, with the exception of the Swedish SweLLex, which is based on CEFR-graded learner essays. Each single word or multiword expression that has been found in textbooks and other language learner material is listed in its base form, i.e. lemmatized, together with an automatically derived part-of-speech tag. For each entry, the resource lists its normalized distribution over the respective CEFR levels as indicated by the learning material. Table 1 shows examples from the English EFLLex.

Other resources aligned to the CEFR that we use in this work are the KELLY lists (Kilgarriff et al., 2014), which exist for nine different languages, including English and Swedish, the Pearson Global Scale of English (GSE) (Pearson, 2017) and the Cambridge English Vocabulary Profile (EVP) (Capel, 2015) vocabulary lists for English.

Two obvious problems that we are facing are the absence of a gold standard for most languages, as well as a data sparseness issue. Indeed, any such list pertaining to natural language must be finite and cannot, by definition, be exhaustive. This may then result in certain expressions being only found at advanced levels, although they could have been introduced much earlier. Furthermore, textbooks may opt to introduce vocabulary of a higher level if necessary for certain tasks, which may give the impression that a word is

Expression	PoS	A1	A2	B1	B2	C1	Total
video	noun	2.47	0.56	34.83	23.80	13.25	18.43
write	verb	934.71	378.34	760.73	536.38	713.33	549.91
empty	adjective	86.49	150.89	65.95	194.80	123.41	156.02
shopping center	noun	0	0	15.58	0	0.82	1.75
dream up	verb	0	0	0	0	0.82	0.23

Table 1: Sample from EFLLex

used earlier, and thus easier to understand, than expected. In such cases, the words in question are often explained further.

In the absence of a gold standard we are using the two mentioned well-established independent English lexicons GSE and EVP as a base for comparison. For our experiments, we use both of them together as gold standard by combinding their scores. As a resource of the same kind, we expect EFLLex to correlate with the gold data.We further assume that the findings for EFLLex also hold for other CEFRLex resources, as they follow the same methodology.

Thus, the purpose of this study is to evaluate the applicability of CEFRLex resources for language learning applications. To this end, we use monolingual lexical resources in English as an external reference, and compare it to the English CEFRLex. Then, with the help of translation candidates from word-aligned parallel corpora, we identify the divergence from CEFR levels in other languages, and evaluate if the chosen features, in combination with other monolingual resources, lead to a better fit. In our experiments, we only consider single words, as multi-word expressions account for only a small share of the lexical entries (see Figure 2 in Section 3.4), and word alignment of multi-word units in parallel corpora is less accurate.

2. Related Work

The KELLY project (KEywords for Language Learning for Young and adults alike) aimed at creating a language learning tool for nine different languages (Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish) (Kilgarriff et al., 2014). To this end, approximately 9,000 keywords were collected for each language, based on their frequency in large corpora. After ordering the list by frequency, it was divided into six equally-sized parts and assigned CEFR levels, from A1 for the most frequent items, to C2 for the least frequent items.

KELLY vocabulary can be seen as "core" vocabulary, i.e. vocabulary that should be known by a prototypical learner of a certain proficiency level. Each resource was also manually translated into all other eight languages. As an added effect of interlinking the lists through translation, it is also possible to identify expressions that occur in all lists ("universal vocabulary"), expressions that occur in most of the lists ("common vocabulary") and expressions that only occur in certain language pairs or only in one single list ("language-specific vocabulary") (Volodina and Kokkinakis, 2012).

For English, two of the most prominent resources which, among other things, link expressions to CEFR levels are the English Vocabulary Profile (EVP) (Cambridge University Press, 2015), and the Global Scale of English (GSE) Teacher Toolkit (Pearson, 2017). While the GSE Teacher Toolkit is freely accessible, EVP requires a (free) subscription.

A possible application for CEFR-graded word lists is, for instance, the readability assessment of texts including visualization of words of different CEFR levels. Projects that employ such a methodology are, inter alia, Duolingo CEFR checker[1] for English and Spanish, Texteval[2] for Swedish and the CEFRLex Lexical Complexity Analyzer[3] for English, Spanish, French and Dutch. Each of these tools highlights words of different CEFR levels in different colors. The first two also incorporate a readability estimation algorithm, which predicts an overall CEFR level for the text, while the latter lets the user select a target CEFR level and highlights all words that belong to a level higher than the chosen one.

In readability assessment research, lexical features have repeatedly shown to be one of the most prominent predictors of readability (Beinborn et al., 2014; François and Fairon, 2012; Heilman et al., 2007; Huang et al., 2011; Pilán et al., 2016; Volodina et al., 2016). It has also been shown that replacing "traditional" frequency-based word lists by CEFRLex-derived resources significantly improves results of automatic essay grading (Pilán et al., 2016).

While tools such as readability assessment of texts can be useful not only to teachers but also to learners, a more learner-targeted application of CEFRLex resources is the automatic generation of exercises, as for example exemplified by the Lärka platform (Alfter et al., 2019) where multiple different exercises such as listening exercises or word guess exercises are automatically generated, or the multilingual particle verb exercise described in (Alfter and Graën, 2019), which connects different language resources, all of which are taking into account the level of proficiency of the learner as well as the estimated proficiency level at which a learner can understand certain words as given by the CEFRLex resource.

Some of our English analyses in this paper have already been conducted for Dutch (Tack et al., 2018), such as frequency effects and word length effects. We go beyond their approach by comparing to a soft gold standard, by comparing several algorithms for calculating the learning level, by suggesting possible changes to the CEFR level, and by profiting from multilingual resources.

[1] https://cefr.duolingo.com/

[2] https://spraakbanken.gu.se/larka/texteval

[3] https://cental.uclouvain.be/cefrlex-demo/analyze

3. Resources

In this work, we use EFLLex for English (Dürlich and François, 2018), FLELex for French (Tack et al., 2016) (see also (François et al., 2014)), and SVALex for Swedish (Francois et al., 2016), all available online.[4] We are aware that CEFRLex resources for Dutch (Tack et al., 2018) and Spanish (François and De Cock, 2018) have been compiled, and that there is ongoing work on creating CEFRLex resources for German and Portuguese as well, but for the scope of this paper, we have chosen to limit ourselves to those three language resources that have officially been released.

It should be noted that there are two different versions of the French CEFRLex resource, differing only in the choice of part-of-speech tagger, and that we have chosen the CRF (Conditional Random Field) version, as this tagger is said to be more accurate (François et al., 2014). It should also be noted that only the French CEFRLex resource covers all six CEFR levels, from A1 to C2. All other CEFRLex resources disregard the C2 level, as it is notoriously difficult to find textbooks pertaining to the highest level of proficiency. At this level, learners have attained near-native proficiency and they have, thus, little need for textbooks.

3.1. Word alignments from a general corpus

The Sparcling corpus (Graën, 2018; Graën et al., 2019) consists of parallel texts in 16 different languages. It comprises the debates of the European Parliament for a time span of 15 years, originally published as Europarl corpus by Koehn (2005) and released in a cleaner version with document-level alignment by Graën et al. (2014). The corpus features alignment on several levels, from documents down to bilingual word alignment for each language pair. Word alignment has been performed with four different word aligners, namely GIZA++ (Och and Ney, 2003), the Berkeley Aligner (Liang et al., 2006), fast_align (Dyer et al., 2013) and efmaral (Östling and Tiedemann, 2016). For the present work, we only used those alignment links that were supported by all four aligners, thus strongly favoring precision over recall.

Based on those word alignments, we derive the conditional probability of a token with lemma λ_s in one language being aligned with a token with lemma λ_t in another language (Graën, 2018, Section 3.2.1). With f_a being the frequency of two lemmas being connected via word alignment of their corresponding tokens, the conditional probability p_a of the target lemma λ_t given the source lemma λ_s is calculated as:

$$p_a(\lambda_t|\lambda_s) = \frac{f_a(\lambda_s, \lambda_t)}{\sum_{\lambda_{t'}} f_a(\lambda_s, \lambda_{t'})}$$

If we also take assigned part-of-speech tags θ into account, this equation extends to:

$$p_a\left((\lambda_t, \theta_t)|(\lambda_s, \theta_s)\right) = \frac{f_a((\lambda_s, \theta_s), (\lambda_t, \theta_t))}{\sum\limits_{(\lambda_{t'}, \theta_{s'})} f_a((\lambda_s, \theta_s), (\lambda_{t'}, \theta_{s'}))}$$

For example, the alignment probability of the French noun 'vaccin' given the English noun 'vaccine' is high (94%). Other correspondences of the English source lemma identified via word alignment are the verb 'vacciner' (to vaccinate), the noun 'vaccination' (vaccination), and, with a single occurrence each, the nouns 'grippe' (influenza) and 'médicament' (medicine/pharmaceutical). The other way round, the alignment probability of English 'vaccine' given French 'vaccin' is also high (91%). Alternative alignments are 'vaccination' and, with a single occurrence, 'inoculate'.

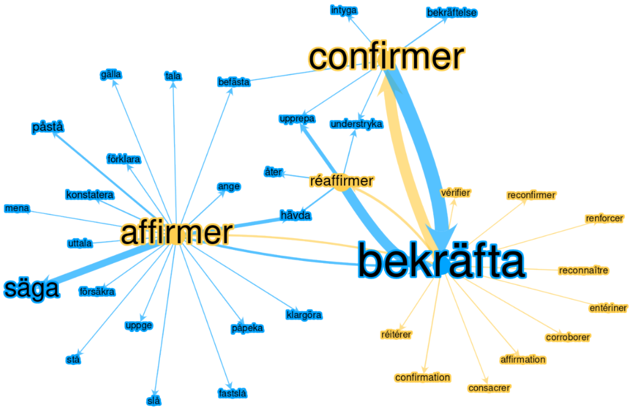


Figure 1: Alignment probabilities for Swedish (blue) and French (yellow) words. The size of the nodes represents corpus frequency and the sizes of the connecting lines relates to conditional alignment probability.

While the lemma 'vaccine' shows a strong alignment and thus translation preference for 'vaccin', and vice versa, other correspondences are not as straightforward. The most frequent alignment of French 'confimer' (to confirm) to Swedish is 'bekräfta', which also holds for the opposite direction. However, 'bekräfta' given 'confirmer' is considerably more probable (93%) than 'confirmer' given 'bekräfta' (70%). Other frequent correspondences of 'bekräfta' are 'réaffirmer' (10%) and 'affirmer' (8%). Figure 1 depicts the alignment probability for those words (Graën and Schneider, 2020).

In case of compounds in one language that correspond to two or more tokens in a second language, the alignment probability is distributed to all constituents of the corresponding expression, e.g. for English 'waste management' and Swedish 'avfallshantering', we see a probability of 50% for 'waste' given 'avfallshantering' and 31% for 'management' given 'avfallshantering'.

For each pair of lexical entries in two CEFRLex resources, we determine the respective conditional alignment probability for both directions from the Sparcling corpus. As we are looking for standard translations, we set a threshold of 25%, below which we ignore alignment probabilities. The alignments of both 'waste' and 'management' with 'avfallshantering' would pass, but for 'bekräfta', we would only accept 'confirmer' with its value of 70%.

3.2. Multilingual core vocabulary

In this paper, we consult the English and Swedish KELLY lists. The KELLY lists can be regarded as core vocabulary, i.e. vocabulary that should be known by a generic learner of a certain level (Kilgarriff et al., 2014). KELLY lists are

[4] https://cental.uclouvain.be/cefrlex/

frequency-based and the assigned CEFR levels directly result from a frequency-ranking of expressions as found in large web corpora.

The English KELLY list was compiled from the UK-Web-as-Corpus (ukWaC) corpus (Ferraresi et al., 2008) and British National Corpus (BNC) (The BNC Consortium, 2007). UkWaC contains over 2 billion words, while BNC contains almost 100 million words. The English KELLY list comprises the 7,549 most frequent lemmas, although they are not evenly distributed across the six CEFR levels.[5]

The Swedish KELLY list was compiled from the Swedish-as-a-Web corpus (SweWaC) containing 114 million words. It comprises the 8,425 most frequent lemmas distributed evenly across the six CEFR levels.[6]

3.3. Independent English lexicons

We regard the English Vocabulary Profile (EVP) and the graded vocabulary part of Pearson's Global Scale of English (GSE) Teacher Toolkit as independent lexical resources. Through their respective web interfaces, one can query words and phrases, and the results include, among other information, assigned CEFR levels.[7] It should be noted that both EVP and GSE list word senses.

Since EFLLex does not distinguish between senses, we have chosen to conflate EVP and GSE senses in such a way as to assume the first level of any polysemous word as the target level.

While EVP seems to be targeting productive knowledge, given that it is mainly based on the Cambridge Learner Corpus (Nicholls, 2003), GSE is slightly more unclear as to whether it targets productive or receptive knowledge.

3.4. Data overview

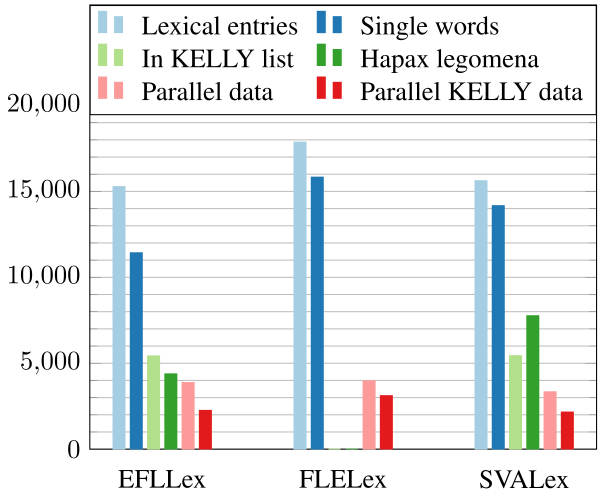


Figure 2: Vocabulary sizes for the three languages

In Figure 2, we show vocabulary sizes from the CEFRLex resources that we use. The French resource, FLELex, unlike the other two, does not include absolute frequencies, hence we cannot detect hapax legomena. Furthermore, there is no KELLY list for French. We do, however, frequently find the translations of French vocabulary entries into the other two languages in their respective KELLY lists ('Parallel KELLY data' in Figure 2). For English and Swedish, there can only be one translation that appears in a KELLY list due to the nonexistence of a French list. This is why we find more translations of French entries in the other languages' lists.

The KELLY lists of English and Swedish comprise less than half of the single-word entries in each language. Between 3,000 and 4,000 entries of each language have parallel correspondences (see Section 3.1).

3.5. CEFRLex combined

In addition to extracting pairwise language combinations as described in Section 3.1, we also created a combined aligned list with entries from all three resources.

To this aim, we start from one language pair, for example French/Swedish, and for each entry we look up possible translations in the third language, English in this case, from the other two resources. Thus, if we start with the French/Swedish list, we retrieve English translations from the aligned English/Swedish and English/French lists. Each entry can have zero, one or multiple translations.

For each translation, we then retrieve its translation probabilities. In case there are multiple translation candidates, we create separate entries. For example, for the French/Swedish noun entry 'question' (French) – 'fråga' (Swedish), two English correspondences are available, namely 'question' and 'issue'. We thus create two aligned entries as (additional information omitted from the example for readability):

PoS	English	French	Swedish
NOUN	question	question	fråga
NOUN	issue	question	fråga

We repeat this process for each of the three paired lists and merge the resulting lists, removing duplicate entries in the process. The resulting list can still contain partial entries (entries with only two languages) that are covered by more complete entries. This is due to the fact that we only perform a single-step translation look-up, and that some translations might only be reachable under certain circumstances. Thus, in a second step, we remove partial entries which are covered by more complete entries.

This results in a final list of 6,077 entries. While the original pairwise files also contain non-lexical part-of-speech entries such as conjunctions, lexical part-of-speech entries (nouns, verbs, adjectives and adverbs) constitute the majority of entries, as listed in Table 2.

Entries can be sparse, i.e. if we do not have a translation candidate in the third language for any given language pair, the entry will only contain the original language pair data. The final combined list contains at least two languages per entry with at least two translation probabilities, up to three languages per entry and $3 \times 2 = 6$ translation probabilities.

[5] A1: 789, A2: 921, B1: 1383, B2: 1107, C1: 948, C2: 2401

[6] Each level from A1 to C1 contains 1404 entries while level C2 contains 1405 entries.

[7] GSE uses a more fine-grained numerical scale from 11 to 89, but also maps this scale onto CEFR levels.

pair	entries	lexical entries
en/fr	4012	3976
en/sv	3329	3298
fr/sv	3350	3319

Table 2: Number of (lexical) entries per language pair

4. Methods

In CEFRLex, each lexical entry (i.e. a pair of lemma and part-of-speech tag) is listed with a distribution of observed frequencies by CEFR level. The frequencies are indicated as relative, normalized, adjusted using dispersion, per-level and per-million-word frequencies (François et al., 2016). The distributions come in different shapes. Figure 3 shows the distribution of 'smör' (butter) in SVALex.[8] We see a peak at B1 level, but the first occurrence of that word in SVALex is at A2.

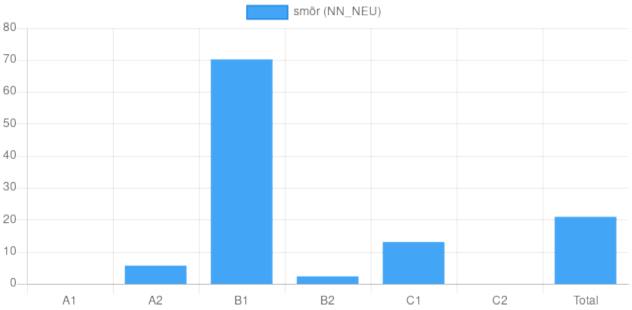


Figure 3: The distribution of the noun 'smör' (butter) in SVALex over CEFR levels from A1 to C1. The numbers represent the expected frequency in one million running words.

The most straightforward strategy to determine the corresponding level for each entry is to go by the first occurrence, in our example that is A2. In some distributions, however, we see a very small number ($\ll 1$) at the first and a considerably larger number at the second occurrence (occasionally enclosing an intermediate level without any reported occurrence). We assume that those might be cases where a word or expression of a higher level has been required for a lower-level text. To account for those cases, we define thresholds of 1%, 5% and 10% of the sum of frequencies over all levels.[9] We refer to the first-occurrence CEFR level as C, to those levels determined by the thresholds of 1%, 5% and 10% as C_1, C_5 and C_{10}, respectively.

In most cases (83%), the resulting levels among all thresholds are the same as the first-occurrence level. Figure 4 shows one of very few cases (4), where all four resulting levels are different. If we go by first occurrence, we would assign the level A2 to the verb 'determine'. With a threshold of 1%, we would skip the A2 frequency (0.29 per million words) and assign the level B1. With the highest threshold of 10%, we finally would skip all lower levels and assign C1 as CEFR level.

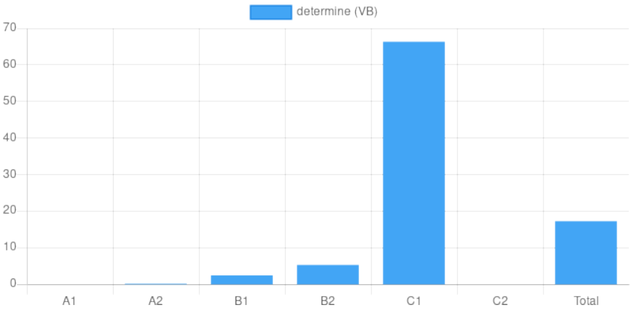


Figure 4: The English verb 'determine' shows comparably low frequencies at lower levels and a peak at C1.

As described in Section 3.1, we identify pairs of lemmas and part-of-speech tags with an alignment probability p_a greater than 25%. For each lexical entry from one of the CEFRLex resources where we find at least one corresponding pair with identical part-of-speech tags, we calculate the minimum (p_{min}), maximum (p_{max}) and average (p_{avg}) of those conditional probabilities in both directions. Minimum and maximum correspond to the directions with a lower and greater probability. For the adjective 'hungry', for instance, we find the Swedish correspondence 'hungrig'. While 92% of the occurrences of 'hungrig' are aligned to 'hungry', 'hungry' is also frequently translated as 'svälta' (to starve) to Swedish, which leaves a 56% probability for 'hungrig'. The minimum is thus 56%, the maximum 96% and the average 74%. In the cases where we find corresponding entries in two languages, we use the average of both minimal, maximal and average values.

Out of 2976 entries that have correspondences in all three languages (with $p_a > 25\%$ in both directions), 406 show the same CEFR level (C) in all three, and 1981 show the same CEFR level in at least two languages. The remaining 995 entries have different levels. The verbs 'work', 'travailler' and 'arbeta', for instance, are all classified as A1, while 'paralyse', 'paralyser' and 'förlama' are classified as C1. On the other hand, we find different levels for 'adventurous', 'aventurier' and 'äventyrlig' (A2, A1 and C1, respectively) or 'linguist', 'linguiste' and 'lingvist' (A2, B1 and C1, respectively).

We calculate the average difference in terms of CEFR levels between our respective target language and the other languages available, and normalize it to the range from -1 to +1, once by dividing it by the maximal distance of 4 levels[10] (δ) and once by using a sigmoidal function (δ_σ), which projects to the same range, but has a more abrupt gradient, thus giving less relative weight to smaller differences.

For each lexical entry (lemma plus part-of-speech tag), we

[8] These charts are generated by the interactive CEFRLex lookup tool located at https://cental.uclouvain.be/cefrlex-demo/search.

[9] The 'total' number, which forms part of each CEFRLex resource (also shown in Figure 3), does not correspond to the sum of frequencies, as each frequency has been normalized to per-million-words over all entries at that level and adjusted by taking dispersion into account. As the total takes all levels into account and the number of observed words per level are different, the numbers do not add up to the 'total' number.

[10] For reasons of comparability, we disregard the near-native level C2, which is only available for French words.

have assembled the following features:

- The four derived CEFR levels (C, C_1, C_5 and C_{10}) mapped to a linear scale (1 = A1, 2 = A2, ...)
- The CEFR level as defined by KELLY (if available)
- A flag whether a word is only seen once in the corpus (hapax legomenon)[11]
- Three values derived from alignment probabilities (p_{min}, p_{max} and p_{avg})
- The number of languages with an alignment probability of more than 25% in both directions
- The number of languages for which we find a corresponding entry in KELLY[12]
- The average difference of corresponding lemmas from the Sparcling corpus (δ) in terms of CEFR levels (normalized to values between -1 and +1)
- The same difference projected to the range -1 to +1 by a sigmoidal function (δ_σ)
- The entry's length in terms of letters
- The entry's frequency from BNC (The BNC Consortium, 2007), both from the entire BNC (100 million words), and just the spontaneous conversation section (4 million words)
- The entry's frequency from ICLE (Granger et al., 2009), a corpus of Learner English, with over 3 million words.

In the absence of a hard gold standard, we rely on some of the best industry efforts and best practices, namely GSE and EVP. These two resources are strongly correlated (the Pearson correlation is 0.85), but there are also differences. Following the logic of ensemble approaches (Dietterich, 1997) or of the four-eye principle, namely that independent systems typically make partly different errors, offer a different perspective and are thus a good base for triangulation, we have decided to predict the sum of GSE and EVP, i.e. their linear combination, to which refer as GSE&EVP in the following.

In order to assess the correlation of EFLLex to GSE and EVP and other correlations, and in order to test out hypothesis that we can further improve EFLLex, we had to restrict our data sets to those cases where we found an entry in both GSE and EVP, and where we obtained a CEFR level. This gives us a data set of 1,571 lemmas. In the smallest lexical resource, KELLY, which mainly reflects core vocabulary, we replaced the frequent null entries by the highest level (C2) in order not to have to restrict our data set further.

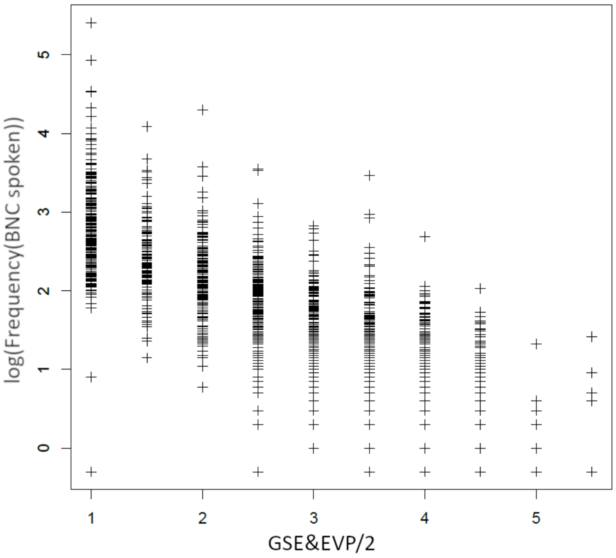


Figure 5: Plot of the correlation between GSE&EVP/2 and log(Frequency(BNC Spoken)). Each dot is a word type.

5. Results

In this section, we report our results. We use EFLLex, and our suggested changes due to multilingual alignment, and we evaluate using GSE and EVP in combination (GSE&EVP, see previous section) as gold standard, and EFLLex and other features as correlated variables and as predictors.

5.1. Correlations

Among the larger set of features that we have tested, we found high correlations between GSE&EVP and the following features: token frequency, word length, C, and our suggested changes to C.

Correlations to individual baseline features are given in Table 3. They confirm and partly extend the findings of (Tack et al., 2018) on Dutch. Concerning frequency, BNC spoken correlates better than the complete BNC, and also better than ICLE, a corpus of Learner English essays. The logarithm of frequency correlates much better, which is in line with psycholinguistic experiments (Smith and Levy, 2013) and Zipf's law. A plot of *log(Frequency(BNC spoken))* vs. GSE&EVP/2 is given in Figure 5. Concerning word length, length (in letters) and its logarithm correlate very similarly. C shows a strong correlation, albeit less well than the trivial feature of frequency from BNC spoken. This fact already indicates that CEFR levels can be approximated further to our assumed gold standard, thus indicating which entries are more reliable and which may need manual verification. C and C_1 correlate almost equally well (C slightly better), increasing the threshold further leads to a decrease in correlation.

In the next step, we test if the model would fit better, if CEFR levels were closer to their counterparts in other languages. We have tested δ and δ_σ using several feature weights, in order to find out: does the correlation increase if we add the suggested correction? What is the approximate optimal weight of the correction?

[11] Not available for FLELex as the absolute frequencies are unknown

[12] As KELLY does not cover French, the value is either 0 or 1 for English and Swedish.

Feature	Pearson Correlation
Frequency(BNC)	-0.1237227
log(Frequency(BNC))	-0.5081583
log(Frequency(BNC spoken))	-0.7820845
log(Frequency(ICLE))	-0.4028432
word length	0.4515713
log(word length)	0.4572295
C	0.7077803
C_1	0.7061353
C_5	0.7027760
C_{10}	0.6802382
KELLY	0.6464615

Table 3: Correlations of individual features to GSE&EVP

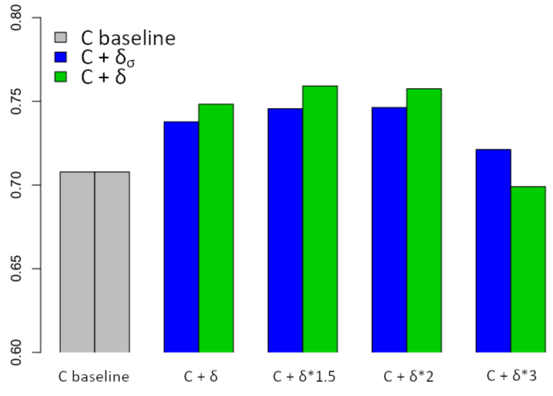


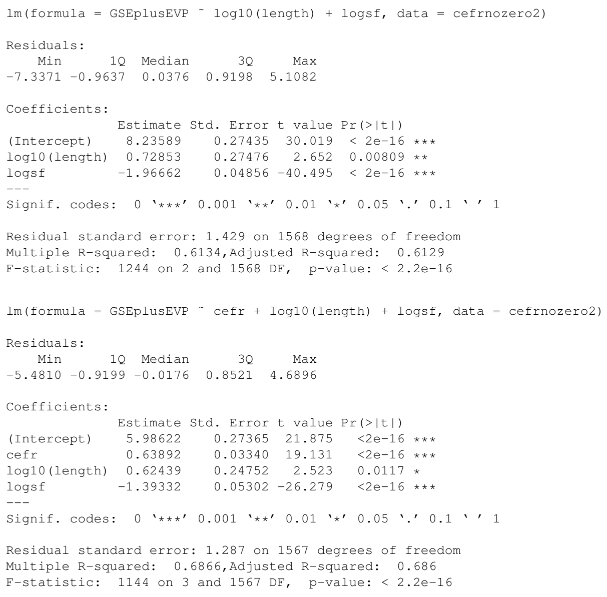
Figure 6: Correlation between CEFR values from GSE&EVP and different combinations of C (CEFR) with the relative CEFR level differences δ (div-lin) and δ_σ (div-curve) from parallel data

Figure 6 shows that δ correlates better than δ_σ, and that optimal weights seem to be around 1.5 or 2.0.[13] The best correlation is 0.759, 0.05 higher than the C baseline. In terms of coefficient of determination (r^2) the proportion of variance increases from $C^2 = 0.708^2 = 50.1\%$ to $0.759^2 = 57.6\%$. Hypothesis 1 has thus been proven.

Correlations can be increased further by adding more features, and adapting the weights of the features. The highest correlations to GSE&EVP reach about 0.85, which is also the correlation between GSE and EVP. A selection of combinations is given in Table 4. The last two lines are baseline feature combinations, indicating that an increase of about 3% can be obtained by our approach.

5.2. Regression Models

Instead of manually tuning feature weights, linear regression models find optimal weights automatically, and distinguish between significant and non-significant features. For example, the flag indicating hapax legomena is not a significant feature.



```
lm(formula = GSEplusEVP ~ log10(length) + logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3371 -0.9637  0.0376  0.9198  5.1082 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.23589    0.27435  30.019  < 2e-16 ***
log10(length)  0.72853    0.27476   2.652  0.00809 ** 
logsf         -1.96662    0.04856 -40.495  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.429 on 1568 degrees of freedom
Multiple R-squared:  0.6134,Adjusted R-squared:  0.6129 
F-statistic:  1244 on 2 and 1568 DF,  p-value: < 2.2e-16


lm(formula = GSEplusEVP ~ cefr + log10(length) + logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4810 -0.9199 -0.0176  0.8521  4.6896 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.98622    0.27365  21.875   <2e-16 ***
cefr           0.63892    0.03340  19.131   <2e-16 ***
log10(length)  0.62439    0.24752   2.523   0.0117 *  
logsf         -1.39332    0.05302 -26.279   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.287 on 1567 degrees of freedom
Multiple R-squared:  0.6866,Adjusted R-squared:  0.686 
F-statistic:  1144 on 3 and 1567 DF,  p-value: < 2.2e-16
```

Figure 7: Baseline Regression models

We discuss five models in the following: First, a low baseline model, predicting GSE&EVP from word length and log of frequency from BNC spoken. Second, an upper baseline which adds C. Third, the model which includes our best correction, $C + \delta \times 1.5$. Fourth, a model which additionally includes $C_1 + \delta \times 1.5$. Fifth, a model which includes all significant features.

First, the low baseline model which uses word length and frequency is given in Figure 7 at the top. It reaches an R^2 value of 61.3%, which can be interpreted as the percentage of the data that is explained by the model.

Second, the model which adds C, but without our suggested CEFR level change, given in Figure 7 at the bottom. Its R^2 is 68.7%.

Third, the lower baseline plus our best-performing CEFR change, $C + \delta \times 1.5$. This factor (δ) is highly significant, as the top half of Figure 8 shows. It reaches an R^2 value of 70.64%

Fourth, our correlation experiments indicated that adding a correction based on C_1, although less well correlated to GSE&EVP than CEFR-based corrections, may help the model. This is indeed the case, as the bottom half of Figure 8 shows. Note that both factors, although highly correlated, stay highly significant.

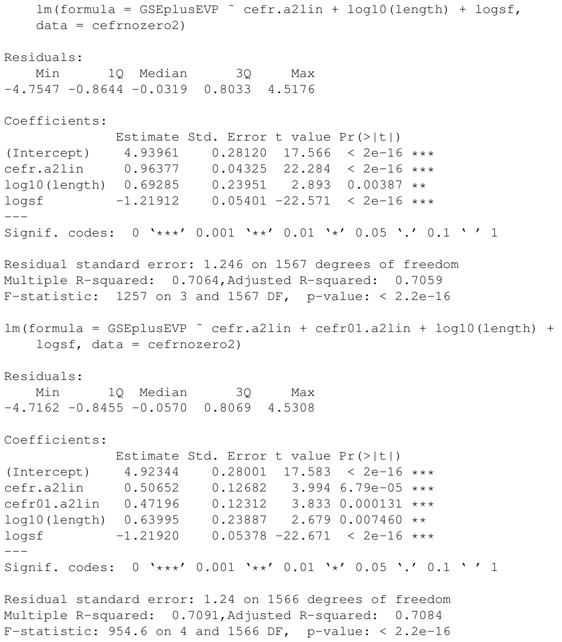
Fifth, the model including all relevant features also adding PoS tags and KELLY information, but neither the hapax legomena flag, nor C_5-based measures, etc. This model reaches R^2 of 72.9%.

Finally, a word on the quality of prediction is due. We consider the output of the fifth model here. The mean of the absolute value of the difference between GSE&EVP/2 to our prediction is 0.46. This means that a prediction is on average off by 0.46 levels. The residuals (for the second and fifth model), given in Figure 9 show a normal distribution, indicating a good model fit. The differences between model 2 and 5 are statistically significant (Welch two sam-

[13] Note that δ and δ_σ are normalized and take values between 0 (no difference found in parallel data) and 1 (a difference of 4 levels, i.e. between A1 and C1).

Features	Pearson Correlation
$C + \delta \times 1.5$	0.7591618
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)})$	0.8352488
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2$	0.8393453
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2 + \log(\mathrm{word\ length})/3.2$	0.8404394
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2 + \log(\mathrm{word\ length})/3.2 + (C_1 + \delta/20)$	0.8405208
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2 + \log(\mathrm{word\ length})/3.2 + (C_1 + \delta/20)$ + hapax/4 − PoS is NOUN/2 + PoS is VERB/7 − PoS is ADJ/3 − PoS is ADV/3	0.8457225
$(C + \delta \times 1.5 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2 + \log(\mathrm{word\ length})/3.2 + (C_1 + \delta/20)$ + hapax/4 − PoS is NOUN/2 + PoS is VERB/7 − PoS is ADJ/3 − PoS is ADV/3 + KELLY/5	0.8517668
$\log(\mathrm{word\ length})/3.2 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2$	0.7828156
$\log(\mathrm{word\ length})/3.2 + C \times 1.2 - \log(\mathrm{f(BNC\ spoken)}) \times 1.2$	0.8210981

Table 4: Correlations of weighted feature combinations to GSE&EVP



```
    lm(formula = GSEplusEVP ˜ cefr.a2lin + log10(length) + logsf,
    data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7547 -0.8644 -0.0319  0.8033  4.5176

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.93961    0.28120  17.566  < 2e-16 ***
cefr.a2lin     0.96377    0.04325  22.284  < 2e-16 ***
log10(length)  0.69285    0.23951   2.893  0.00387 **
logsf         -1.21912    0.05401 -22.571  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.246 on 1567 degrees of freedom
Multiple R-squared:  0.7064,Adjusted R-squared:  0.7059
F-statistic:  1257 on 3 and 1567 DF,  p-value: < 2.2e-16

lm(formula = GSEplusEVP ˜ cefr.a2lin + cefr01.a2lin + log10(length) +
    logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7162 -0.8455 -0.0570  0.8069  4.5308

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.92344    0.28001  17.583  < 2e-16 ***
cefr.a2lin     0.50652    0.12682   3.994 6.79e-05 ***
cefr01.a2lin   0.47196    0.12312   3.833 0.000131 ***
log10(length)  0.63995    0.23887   2.679 0.007460 **
logsf         -1.21920    0.05378 -22.671  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.24 on 1566 degrees of freedom
Multiple R-squared:  0.7091,Adjusted R-squared:  0.7084
F-statistic: 954.6 on 4 and 1566 DF,  p-value: < 2.2e-16
```

Figure 8: Central Factor Regression models

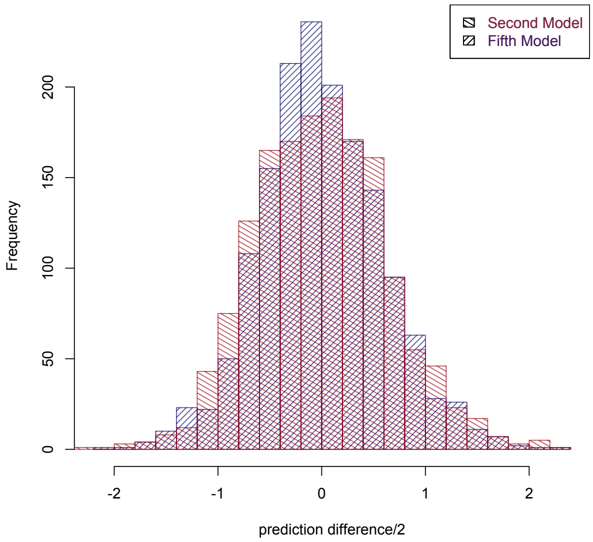


Figure 9: Residuals of the second and fifth model

	Upper Base=2nd Model						5th Model					
	0	1	2	3	4	5	0	1	2	3	4	5
1	5	80	155	1	2	0	2	143	95	3	0	0
2	1	34	354	144	13	1	0	51	314	167	15	0
3	0	0	46	153	20	2	0	0	35	167	19	0
4	0	1	40	202	237	59	0	0	23	210	253	53
5	0	0	0	0	6	3	0	0	0	0	3	6
6	0	0	0	4	1	7	0	0	0	1	3	8

Table 5: Confusion Matrix of Upper Baseline (second model) vs. fifth model. Predicted is the horizontal, actual in the vertical axis. While the actual range is between 1 (A1) and 6 (C2), the models predict (rounded) values between 0 (below A1) and 5 (C1).

ple t-test, $p = 0.0004$), tested on the residuals of the second and fifth model, see Figure 9. This means that the residuals are significantly smaller on the fifth model than on the second model.

Also confusion matrices confirm the improvement. If we round the prediction of the linear models to the nearest integer, we obtain the confusion matrices given in Figure 5. The upper baseline predicts 827 words correctly (out of 1,571), the fifth model 883.

The upper baseline model (the second model) is off by 0.51 levels on average. C on it own is off by 0.83 levels on average, partly due to the fact that C is 0.6 levels higher than GSE&EVP/2. A model predicting GSE&EVP/2 from C only is off by 0.65 levels.

6. Discussion

The reduction of prediction difference from 0.51 levels off (upper baseline = second model) to 0.46 levels off (fifth model) seems modest. But we need to bear in mind that we are dealing with several ceiling effects. First and foremost, word length and word frequency (particularly from BNC spoken) are very strong, and partly orthogonal predictors. With a correlation of -0.78 to BNC spoken, word frequency stays the strongest predictor in all models. In the third model, our C-based correction almost reaches the weight of frequency from BNC spoken. At a correlation of 0.71, also C itself is a strong predictor. It is remarkable that our suggestions can lead to a further approximation to our assumed gold standard of GSE and EVP in combination.

The fact that also GSE and EVP, although best efforts and the achievements of best practice from several decades of teaching experience, cannot be regarded as a clear gold standard, but maximally a good proxy to one, is a major limitation of our study.

7. Conclusion and Future Work

In this study, we examined the correlation between the English CEFRLex resource and a soft gold standard. We have found that the CEFRLex-derived levels are highly congruent with our gold standard. The observed deviations are to be expected, as the combined scores of GSE&EVP seem to model productive knowledge, while CEFRLex reflects receptive knowledge; vocabulary is expected to first be understood receptively before it is used productively.

In the future, we would like to include evaluations with French resources, include psycholinguistic variables such as age-of-acquisition, imageability, concreteness, etc., and add eye-tracking reading times. Furthermore, given that our study suggests a good correlation of CEFR levels across three languages, it would be interesting to try and project CEFR levels from these resources to other, possibly under-resourced, languages for which there are no CEFRLex resources.

All features that we calculated, our derived best-fit CEFR level and the multilingual combined entries from the three CEFRLex resources are available at http://pub.cl.uzh.ch/purl/multiCEFRLex.

8. Acknowledgments

This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1_184212 through the project "From parallel corpora to multilingual exercises: Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications", conducted at Pompeu Fabra University in Barcelona (with Grael, Grup de Recerca en Aprenentatge i Ensenyament de Llengües, and at the University of Gothenburg (with Språkbanken Text).

This research report has made use of the English Vocabulary Profile (EVP). EVP is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile program, which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information.

9. Bibliographical References

Alfter, D. and Graën, J. (2019). Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

Alfter, D., Borin, L., Pilán, I., Tiedemann, T. L., and Volodina, E. (2019). Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. In *Selected papers from the CLARIN Annual Conference 2018*, pages 1–14. Linköping University Electronic Press.

Beinborn, L., Zesch, T., and Gurevych, I. (2014). Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Cambridge University Press. (2015). English Vocabulary Profile. https://www.englishprofile.org/wordlists. Accessed: 2019-11-11.

Capel, A. (2015). The English Vocabulary Profile. In Julia Harrison et al., editors, *English Profile in Practice*, chapter 2, pages 9–27. Cambridge University Press.

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

Dietterich, T. G. (1997). Machine learning research: Four current directions. *AI Magazine*, 18(4):97–136.

Dürlich, L. and François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *11th International Conference on Language Resources and Evaluation*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 644–649. Association for Computational Linguistics (ACL).

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC) – Can we beat Google?*, pages 47–54.

François, T. and De Cock, B. (2018). ELELex: a CEFR-graded lexical resource for spanish as a foreign language. In *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*.

François, T. and Fairon, C. (2012). An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics.

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)*, pages 3766–3773.

François, T., Volodina, E., Pilán, I., 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)*, pages 213–219.

Graën, J. and Schneider, G. (2020). Exploiting multiparallel corpora as measure for semantic relatedness to support language learners. In David Levey, editor, *Strategies and Analyses of Language and Communication in Multilingual and International Contexts*. Cambridge Scholars Publishing.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 222–227. Stiftung Universität Hildesheim.

Graën, J., Kew, T., Shaitarova, A., and Volk, M. (2019). Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection. In Peter Bański, et al., editors, *Challenges in the Management of Large Corpora (CMLC)*. Leibniz-Institut für Deutsche Sprache, 6.

Graën, J. (2018). *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning*. Ph.D. thesis, University of Zurich.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International corpus of learner English*, volume 2. UCL, Presses Universitaires de Louvain.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Huang, Y.-T., Chang, H.-P., Sun, Y., and Chen, M. C. (2011). A robust estimation scheme of reading difficulty for second language learners. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*, pages 58–62. IEEE.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86. Asia-Pacific Association for Machine Translation (AAMT).

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 104–111. Association for Computational Linguistics (ACL).

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*. Cambridge University Press.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Pearson. (2017). GSE Teacher Toolkit. `https://www.english.com/gse/teacher-toolkit/user/lo`. Accessed: 2019-11-11.

Pilán, I., Alfter, D., and Volodina, E. (2016). Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302 – 319.

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 the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

The BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). `http://www.natcorp.ox.ac.uk/`. Distributed by Bodleian Libraries, University of Oxford.

Volodina, E. and Kokkinakis, S. J. (2012). Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 1040–1046.

Volodina, E., Pilán, I., and Alfter, D. (2016). Classification of Swedish learner essays by CEFR levels. *CALL communities and culture–short papers from EUROCALL*, 2016:456–461.

10. Language Resource References

Francois, Thomas and Volodina, Elena and Pilán, Ildikó and Tack, Anaïs. (2016). *SVALex*. ISLRN 854-377-992-687-3.

Anaïs Tack and Thomas Francois and Anne-Laure Ligozat and Cédrick Fairon. (2016). *FLELex*. ISLRN 742-240-876-017-1.