Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-02838-4 OPEN

CEFR vocabulary level as a predictor of user interest in English Wiktionary entries

Robert Lew [1]✉ & Sascha Wolfer [2]

This contribution explores the relationship between the English CEFR (Common European Framework of Reference for Languages) vocabulary levels and user interest in English Wiktionary entries. User interest was operationalized through the number of views of these entries in Wikimedia server logs covering a period of four years (2019–2022). Our findings reveal a significant relationship between CEFR levels and user interest: entries classified at lower CEFR levels tend to attract more views, which suggests a greater user interest in more basic vocabulary. A multiple regression model controlling for other known or potential factors affecting interest: corpus frequency, polysemy, word prevalence, and age of acquisition confirmed that lower CEFR levels attract significantly more views even after taking into account the other predictors. These findings highlight the importance of CEFR levels in predicting which words users are likely to look up, with implications for lexicography and the development of language learning materials.

[1] Faculty of English, Adam Mickiewicz University, Poznań, Poland. [2] Leibniz-Institut für Deutsche Sprache, Mannheim, Germany. ✉email: rlew@amu.edu.pl

Introduction

CEFR: goals and principles. The CEFR acronym stands for the Common European Framework of Reference for Languages. It is a framework developed by the Council of Europe and is widely used for describing language proficiency levels and learning outcomes. In this section, we shall sketch the goals and principles of CEFR, before focusing specifically on the relationship between CEFR vocabulary levels and users' interest in words as reflected in dictionary searches.

The CEFR serves as an important resource in the area of language education, offering a unified basis for outlining language learning objectives and outcomes (Council of Europe, 2001). This framework significantly aids curriculum planners and language teachers in determining the desired proficiency levels for learners, establishing clear objectives, content, and methods. What sets the CEFR apart is its emphasis on communicative competence, aiming to enhance learners' ability to effectively engage in real-life situations through the practical application of language skills in speaking, listening, reading, and writing (Council of Europe, 2023). This approach shifts the focus from theoretical knowledge to the practical usage of language, fostering real communicative proficiency.

The CEFR framework is widely used worldwide, more so than any other framework (Foley, 2019, 29). Nevertheless, some national alternatives do exist, such as the CES in China (Ministry of Education of the People's Republic of China, 2018); and there are also national adaptations or extensions of the CEFR, such as the CEFR-J in Japan (Negishi et al., 2013) or FRELE-TH in Thailand (Hiranburana et al., 2017). However, the fact that most of these use CEFR at least as a starting point testifies to the wide applicability and popularity of the CEFR framework.

As a descriptive framework, CEFR characterizes the capabilities of learners at various stages of language proficiency through "can-do statements." These statements provide a clear depiction of what learners are capable of at each level, rather than dictating specific language content or curricular structures. The utility of the CEFR extends beyond just learner assessment; it plays a crucial role in the design of language curricula, teaching programs, learning materials, and assessment tools. Moreover, it offers a standard for comparing curricula, textbooks, courses, and examinations across different languages (Council of Europe, 2018). In line with contemporary educational priorities, the CEFR principles also underscore the importance of developing plurilingual and intercultural competence. This approach champions the acquisition of multiple languages and the understanding of diverse cultures and perspectives as central objectives of language learning (Çelik, 2013). This holistic view of language education puts emphasis on communicative abilities but also encourages a broader cultural understanding and adaptability among learners.

The CEFR is closely linked to the European Language Portfolio (ELP), an instrument based on the CEFR (Little, 2009). The ELP allows individual learners to reflect upon their learning, record their progress, and self-assess their language proficiency using the CEFR descriptors. As mentioned above, these descriptors are formulated in terms of "can-do statements" appropriate to the language abilities of learners at each level. These statements outline specific tasks and activities that learners are expected to be able to perform, helping learners set realistic goals and track their progress. At the same time, CEFR recognizes that language learning is a lifelong process. It encourages learners to continue improving their language skills beyond the classroom and provides a foundation for further language development and self-assessment (Council of Europe, 2020, 244).

The CEFR framework is not without problems. Critics have pointed out the insufficient nuance and rigidity of the CEFR levels, which are not in principle sensitive to context of use or learning context (for CEFR levels see below; Krumm, 2007, Widdowson, 2015). Further, the CEFR uses speakers as the benchmark (Widdowson, 2015) and has limited usefulness in standardized language testing (Weir, 2005). Alderson (2007) also notes the vague language used in the framework that is difficult to translate into actual practice.

More pertinently to the goals of this study, the CEFR defines six common reference levels: A1, A2, B1, B2, C1, C2 (Council of Europe, 2020, 36–37), where A1 is the most elementary ("breakthrough") and C2 most advanced ("mastery"). These levels describe the learner's proficiency in reception, production, interaction, and other language competences. These reference levels provide a basis for comparing language curricula, textbooks, courses, and exams.

A crucial component of any language is its vocabulary stock, forming the essential "building blocks" of texts. The CEFR levels can be, and have been, used to grade vocabulary items, specifying the typical learner level at which specific words (or, sometimes, their specific senses or uses) are appropriately known. Unfortunately—and perhaps surprisingly given the fundamental role of vocabulary in communication—the CEFR initiative within the Council of Europe has stopped short of constructing any "official" inventories of vocabulary levels. Those that do exist appear to be independent efforts, often in the context of projects involving the creation of learning materials, and many of them commercial in nature, which makes it difficult to source CEFR wordlists for research purposes.

CEFR levels and interest in vocabulary items. Given that one of the aims of the CEFR is to assign measures of relevance to vocabulary items, it is worth examining how well its grading system aligns with observable behavior. A possible metric to use is the frequency of online dictionary searches. Essentially, if CEFR grading truly reflects a word's importance, then lower-level vocabulary items should find more interest from online dictionary users. As it happens, there is only one online dictionary for the English language—the English Wiktionary—that is well-suited to this purpose, because it is freely available, has exceptionally extensive coverage of words (far more than any available alternative; at the time of this writing, January 2024, about 8 million entries overall, of which about 1 million are English entries), and—crucially—has data on usage freely available for download. These statistics of users visiting Wiktionary pages holding specific dictionary entries can be adopted as a useful operationalization of user interest in English vocabulary items.

Turning now to the possible source of CEFR grading for English, there exist at least two CEFR-graded vocabulary lists that appear to enjoy a degree of prestige while being of non-trivial size, that is on the order of at least thousands of items. These are the *English Vocabulary Profile* (Capel, 2012, 2015, Cambridge University Press, 2015) and the *Oxford Learner's Word Lists* (Oxford University Press Oxford Learner's Word Lists (2023)). Due to copyright considerations, the *English Vocabulary Profile* was adopted as a source of CEFR levels for English words in this research.

Dictionary logs as a source of data. Logs of dictionary visits have been studied before. One motivation behind such studies has been to examine user behavior (Müller-Spitzer et al., 2015), mostly with an eye to improving dictionaries to better serve their user (Lemnitzer, 2001, Bergenholtz and Johnson, 2005, Lorentzen and Theilgaard, 2012, Trap-Jensen et al., 2014). Another reason for examining dictionary logs has been to explore the relationships between user visits and lexical frequency. This relationship

was of relevance in assessing the appropriateness of corpus-based methods in dictionary-making. In particular, a series of studies sought to establish whether corpus frequency—i.e., the relative number of occurrences of a word in a large collection of texts—was a sensible guide in choosing headwords to include in a dictionary. If it could be shown that corpus frequency can predict what users look up in a dictionary, and if this relationship turns out to be positive, then lexicographic work can be optimized to prioritize high-frequency items. This line of research initially brought disappointing findings, with several early studies only finding a weak positive relationship for a few thousand most frequent items, but not much of a pattern beyond that threshold (De Schryver and Joffe, 2004, De Schryver et al., 2006, Verlinde and Binon, 2010). However, further research using more refined methods showed that the effect in the lower frequency ranges was masked by a long tail of very low-frequency items, some of which are nevertheless sometimes looked up (Koplenig et al., 2014a, Müller-Spitzer et al., 2015). Koplenig et al. (2014a) demonstrated that dictionary views are positively related to corpus frequency, while Müller-Spitzer et al. (2015) also found a positive effect of polysemic status (i.e., the word having more than one sense in the dictionary).

Two further possible factors, albeit having less impact, were recently identified by Lew and Wolfer (2024): word prevalence and age of acquisition. This last study, also using logs of Wiktionary views but for a shorter period, suggests that user interest in English entries may be affected by a number of lexical factors, namely (in decreasing measure): (1) a word's lexical frequency; (2) its polysemy status (whether monosemous or polysemous); (3) word prevalence; and (4) age of acquisition. These findings should also be taken into account in investigating the relationship between a word's CEFR level and its look-up frequency, and we intend to do just that in the present study.

As we have previously presented the concept of CEFR levels, the reader also deserves some background on the other four factors as identified in related research discussed above.

The first of these is *lexical frequency*, which is the frequency of a lemma (or, sometimes, a word-form) in a corpus (=a large digital collection) of texts. Frequency can simply be a raw count of how often a word appears in a corpus, although it is often expressed as a standardized frequency per one million tokens to make it independent of corpus size. As discussed earlier, most studies have found that words that are higher in frequency tend to attract more user interest (Koplenig et al., 2014b, Müller-Spitzer et al., 2015, De Schryver et al., 2019).

Polysemy is a property of a lexical item having more than one sense, or distinct meanings (see e.g., Van der Meer 2004). Müller-Spitzer et al. (2015) found that polysemous words tend to attract more views.

The *prevalence* of a word indicates how widely it is known among speakers (Weizman and Snow, 2001, Longobardi et al., 2015). In simpler terms, a word that has high prevalence is known to a large proportion of speakers. One study (Lew and Wolfer, 2024) found a weak negative effect of a word's prevalence on the tendency to look it up in a dictionary.

Finally, *age of acquisition* (Garlock et al., 2001, Juhasz, 2005, Kuperman et al., 2012) refers to the typical age at which a given word is learned in a naturalistic setting. At least one study (Lew and Wolfer, 2024) found that words acquired late in life tend to attract more views than those learned in early life.

Reflecting on the underlying nature of the factors discussed above, we might write that lexical frequency indicates how common a word is in a large collection of texts, which in turn corresponds to the frequency with which people come across a word. Polysemy status describes, as it were, semantic versatility (as seen by creators of the entry for the word). Age of acquisition is a (recalled and reported) typical age at which the word is learned, reflecting a certain developmental sequence that in turn tells us about how important a word is to a developing human speaker at a particular stage. Prevalence is about how widely a given word is known in the population. These measures are not independent. Conversely, it is reasonable to expect that words with many senses, words learned at an early age, and words known to many people are more likely to be more frequent on average, other things being equal. In modeling the relationship between these variables and user interest (as indicated by entry views), these mutual interrelationships should be taken into account.

CEFR vocabulary levels provide information about words that is qualitatively different: these levels reflect expert judgements about a word's usefulness to language learners at a particular stage. As such, CEFR levels have a pedagogical purpose. In what follows, we shall attempt to assess whether CEFR vocabulary levels correspond to the degree of user interest in specific words, as measured by dictionary consultation.

Study

Aims. We derive the following research questions and hypotheses from our introductory remarks.

Question 1: Does the CEFR vocabulary level of an English word (as defined in the English Vocabulary Profile) carry predictive power with respect to how often an entry for this word is viewed in the English Wiktionary?

We especially focus on the CEFR levels A1 to B2 and, to a lesser extent on the conflated C1/2 levels. We take the views in the English Wiktionary as a quantifiable measure of user interest in vocabulary items.

Hypothesis 1: Our first hypothesis is a directed one: lower-level vocabulary items should find more interest from online dictionary users. In terms of views, we expect a hierarchy of A1 > A2 > B1 > B2 > C1/2.

Question 2: If Hypothesis 1 holds, we are further interested in whether the CEFR level still carries independent predictive power as soon as the effects of other lexical variables (e.g., word frequency) are controlled for.

Hypothesis 2: The CEFR level of an English word carries independent predictive power over other lexical variables. We do not have a strong hypothesis as to where the CEFR level fits into the hierarchy of predictors. However, we assume that at least the established effect of corpus frequency is stronger than that of the CEFR level.

Method

Data sources. Data on English Wiktionary views were derived from daily server logs covering a period of four years from 2019-01-01 to 2022-12-31. These statistics are provided by the Wikimedia Foundation and are freely downloadable, but the volume of the original data is huge, given that the data include page visits from all Wikimedia projects (Wikipedia, Wiktionary, Wikidata, Wikimedia, Wikinews, etc.) and needed to be filtered to select the data specific to the English Wiktionary.

We sourced the *CEFR levels* for English lexis from the English Vocabulary Profile (Capel, 2015). Since there was no way of knowing which exact sense users sought in the Wiktionary, we always used the lowest CEFR level if an item happened to have different CEFR levels for different senses. This seemed to make the most sense, given that these were the most frequent and essential senses and thus likely to drive the larger part of dictionary consultations. In any case, the majority of items only had single CEFR levels, so this was not a major issue.

Table 1 Coverage of items from the CEFR dataset in the Wiktionary, broken down by CEFR Level.

CEFR Level	Total	In Wiki	Coverage
A1	645	575	89%
A2	1081	860	80%
B1	1881	1367	73%
B2	2562	1710	67%
C1	1506	993	66%
C2	2084	1003	48%

Further, *lexical frequency* information was drawn from the SUBTLEX-US corpus (as described in Brysbaert and New, 2009). To obtain information on *polysemy*, we extracted the number of senses for each entry word in the English Wiktionary. The custom extraction function written in R (R Core Team, 2023) accesses the edit page of each article, and is available upon request. *Age-of-acquisition* (AoA) ratings were extracted from the supplementary material attached to Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). Finally, *word prevalence* values derive from the supplementary material published alongside Brysbaert et al. (2019). Both AoA and prevalence were assessed by human participants in very large language surveys. The methodology is described in detail in the respective publications.

Modeling. To verify if, and to what extent, the CEFR levels of lexical items predicted the number of views of their entries in the Wiktionary, we ran a regression model (after checking model assumptions) predicting logged views on the CEFR level of the same lexical item. Because the CEFR scale is meant to progress in stepwise fashion from A1 all the way to C2, we used the contrast coding that gave justice to this idea, namely backward-difference coded contrasts, using the R package codingMatrices (Venables, 2023). Due to the poor representation of the highest C1 and C2 levels in the CEFR resource—each one holding fewer vocabulary items than either of the B levels—we conflated these two levels into a single combined level C.

As the distribution of Wiktionary look-up data is skewed towards low values (a large proportion of entries are looked up only occasionally), we log-transformed the *views* variable, achieving a near-normal distribution for the purpose of model computation. Similarly, we used log-transformation on the standardized corpus frequency variable. This ensured that the distribution of the residuals of the linear model was near-normal. Further, we also standardized all continuous predictors (i.e.,: age of acquisition, log frequency, and prevalence) to *z*-scores by subtracting the mean from each value and then dividing by the corresponding standard deviation. This operation brings the values of the predictors into the same range with a mean value of 0 and a standard deviation of 1, making linear regression model estimates comparable.

Results

Words with CEFR levels versus words in the Wiktionary. To see whether CEFR-level information was related to user interest in the associated dictionary entries, we crossed it with information on the identity and popularity of all existing English Wiktionary entries drawn from Wikimedia logs.

The CEFR resource yielded nearly ten thousand items, both single- and multi-words, each with an assigned CEFR level. Of these items in the CEFR dataset, 67 per cent had an equivalent English Wiktionary entry, but this varied by CEFR level, with details as given in Table 1. It will be seen that the coverage declines with increasing CEFR level, from nearly 90 per cent for the most basic A1 level, down to only about half at the highest CEFR level of C2.

Table 2 Parameter estimates of a regression model predicting logged views from CEFR levels.

Term	Estimate	Std. Error	*t*-value
Intercept	10.251	0.012	822.10
Level A2-A1	−0.728	0.049	−14.90
Level B1-A2	−0.318	0.040	−8.04
Level B2-B1	−0.341	0.033	−10.36
Level C-B2	−0.329	0.030	−10.99

CEFR level and views. In order to see to what extent the CEFR levels of lexical items predicted the number of views of their entries in the Wiktionary, we regressed the logged views on the CEFR level of the corresponding lexical item. The model's R^2 was 0.235 (adjusted R^2: 0.234), and the coefficient estimates are given in Table 2. It will be seen that all two-way differences between consecutive CEFR levels are significant, meaning that the step from one to the other makes a significant contribution to predicting user views. The direction of the relationship is such that words assigned a higher (i.e., more advanced) CEFR levels tend to attract fewer views. Conversely, more basic vocabulary items tend to get more views.

The outcome of this model is summarized graphically in Fig. 1, which also includes a beeswarm-plot representation of the complete underlying data.

Thus far, we have explored the relationship between CEFR level and the number of views of the corresponding Wiktionary entry. We will now expand our model to include more information for the English items: specifically, the four lexical factors outlined earlier in this paper: (1) lexical frequency; (2) polysemy; (3) prevalence; and (4) age-of-acquisition. In this part of the analysis, we only kept items that had all these pieces of information.

Refining the model: CEFR + lexical variables. Having combined the CEFR information with the other four lexical predictors, we fitted a multiple regression model with Wiktionary views (logged) as the outcome variable and standardized predictors. The model's R^2 was 0.503 (adjusted R^2: 0.502) and parameter estimates are given in Table 3.

The CEFR-level effect in this more complex multiple regression model was significant, except in the pairwise difference B2-B1, which was marginally significant, while C-B2 was just shy of marginal significance (at $p = 0.105$). This suggests that CEFR is a useful predictor of user interest even once the other predictors are controlled for. Frequency, polysemy, prevalence, and age-of-acquisition were all highly significant and effects pointed in the previously suggested directions (note the positive versus negative signs in the Estimate column of Table 3). Because one would expect the predictors to be correlated, we checked for collinearity using the VIF and the $\mathrm{GVIF}^{1/(2\cdot \mathrm{df})}$. The magnitude of variance inflation appeared to be in safe territory for all predictors (with the highest VIF values being 2.33 for the VIF and 1.53 for the $\mathrm{GVIF}^{1/(2\cdot \mathrm{df})}$). The CEFR effects estimated by this model are summarized graphically in Fig. 2, showing predicted mean views with 95% Confidence Intervals.

Relative importance of CEFR versus other predictors. To estimate the relative importance of the predictors, we computed the

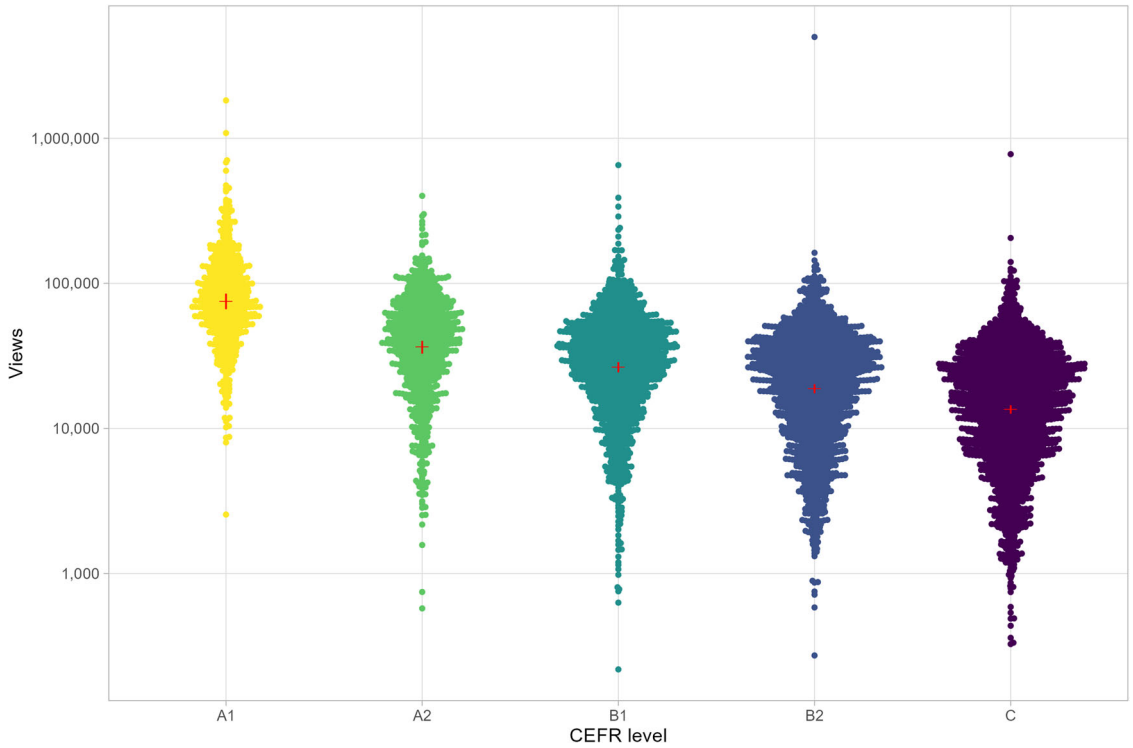

Fig. 1 Views (log scale) by CEFR level. Levels C1 and C2 have been conflated as C. Each colored dot marks one word. Predicted values of Wiktionary views from a linear model with CEFR level as predictor, with 99.9% Confidence Intervals in red.

Table 3 Parameter estimates of a regression model predicting logged views from CEFR level, frequency, polysemy, prevalence, and age of acquisition.

Term	Estimate	Std. Error	*t*-value	*p*-level
Intercept	9.629	0.032	297.48	<0.001***
Frequency	0.572	0.013	43.16	<0.001***
Polysemy TRUE	0.622	0.034	18.50	<0.001***
Age-of-acquisition	0.037	0.012	2.98	0.003***
Prevalence	−0.033	0.009	−3.52	<0.001***
Level A2-A1	−0.158	0.039	−4.06	<0.001***
Level B1-A2	−0.067	0.031	−2.18	0.029**
Level B2-B1	−0.051	0.026	−1.94	0.052*
Level C-B2	−0.040	0.025	−1.62	0.105

Pairwise differences between consecutive levels in the model were significant except the difference between levels B2 and C (conflated). The higher the level, the lower the views. *P*-level ranges: ***<0.01; **<0.05; *<0.1.

change in adjusted R^2 after dropping each predictor from the model, as given in Table 4. The adjusted R^2 of the complete model was 0.502. A reduction in adjusted R^2 when a given predictor is dropped from the model represents the loss of explanatory power of the model, and thus the relative importance of that predictor. It will be seen from Table 4 that CEFR level comes out as the third most important predictor, following lexical frequency and polysemy, but ahead of prevalence and age-of-acquisition.

Discussion and implications

The English Vocabulary Profile CEFR-graded vocabulary exhibits marked variation in terms of its coverage in the English Wiktionary. The coverage is nearly complete (90 percent) for the most basic A1 level, but declines with increasing CEFR level, so that at the highest CEFR level of C2, only about every second vocabulary items is covered. This is to be expected, given that the lower levels of the CEFR scale are designed to cover the most basic vocabulary items, while the vocabulary range at CEFR level C2 is in principle open-ended. In addition, items at the higher CEFR levels in the English Vocabulary Profile often contain lexicographic placeholders (such as *sb*, *sth*) and special symbols such as slashes. For the current contribution, we did not try to resolve these but went with exact matches between CEFR item and Wiktionary page/entry title.

Our study shows that CEFR level is a useful predictor of user interest in English lexical items. The relationship is such that words from the lower (more basic) levels are looked up more in the English Wiktionary, and that in the simple regression model this relationship holds for each individual increment between CEFR levels (except between C1 and C2, for which the data are very incomplete, so these have been conflated as C). Adding the four further lexical predictors—that is corpus frequency, polysemic status, age-of-acquisition, and prevalence—improves the model in terms of its predictive power (the coefficient of determination R^2 doubles), which suggests that these lexical factors also hold useful information that translates into increments in user interest. Nevertheless, CEFR remains a significant factor also in the multiple regression model. Specifically, in this model, the increments from A1 to A2 and from A2 to B1 were significant, B1 to B2 marginally significant, and B2 to C came very close to marginally significant. The inclusion of these other lexical factors ensures that CEFR level is not merely a proxy for another factor, notably lexical frequency. Thus, CEFR level contributes additional information on how useful a lexical item is to Wiktionary users, over and above that carried by the frequency of the item, its age-of-acquisition, prevalence, and polysemy status. Comparing the relative importance of the five variables in terms of how they affect user views, it will be seen from Table 4 that CEFR level ranks as the third most important predictor, following lexical frequency and polysemy, but ahead of prevalence and age-of-acquisition.

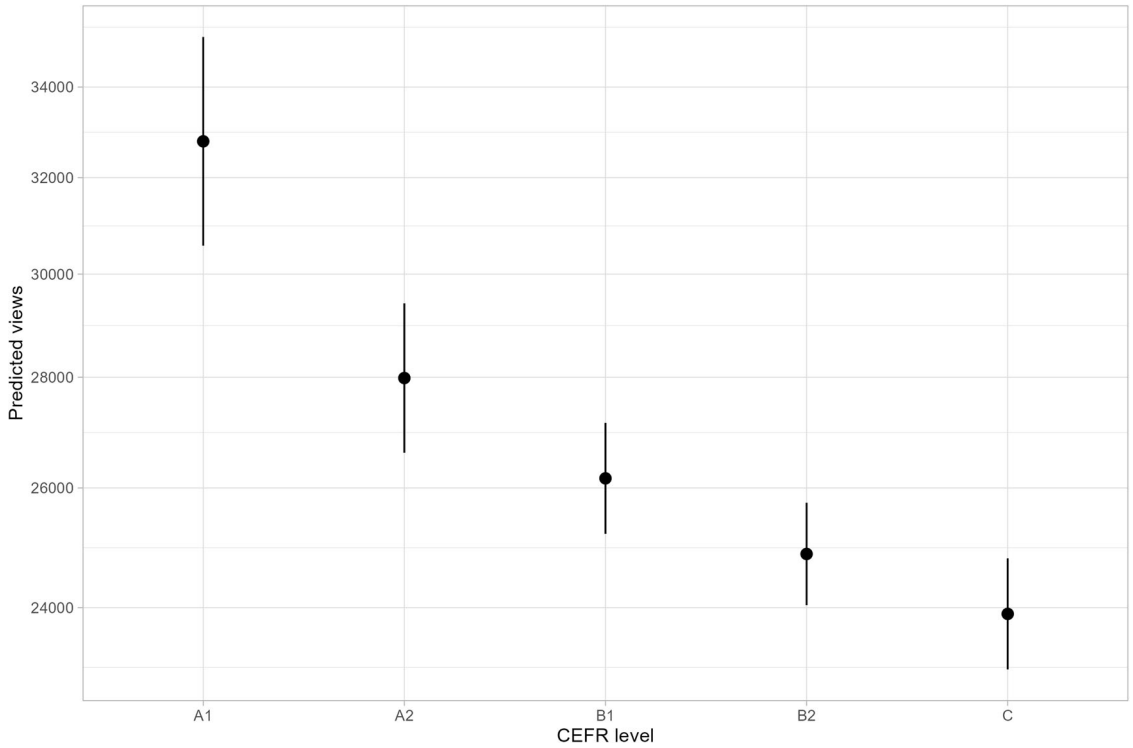

Fig. 2 Effect of CEFR Level on Views in the English Wiktionary after correcting for effects of frequency, polysemy, prevalence, and age of acquisition. Shown are predicted mean views with 95% Confidence Intervals. Levels C1 and C2 have been conflated as C. Lower CEFR levels are viewed more often.

Table 4 Adjusted R^2 and reduction in adjusted R^2 after dropping individual predictors, ordered by decreasing importance.

Predictor	R^2_{adj}	Delta R^2_{adj}
Lexical Frequency	0.341	−0.161
Polysemy	0.472	−0.0295
CEFR Level	0.498	−0.00423
Prevalence	0.501	−0.000985
Age of Acquisition	0.501	−0.000678

In conclusion, the two analyses presented here suggest the presence of the effect of CEFR when considered as a sole predictor, but also in combination with four other lexical factors in a multiple regression model. All in all, the idea to include CEFR information in dictionaries, particularly at the sense level, is a promising avenue of research which could benefit both lexicography and the field of second-language learning.

Limitations and future work

CEFR levels are primarily designed with language learners in mind, but the Wikimedia page view data do not (and cannot) include any information on the personal characteristics of page visitors. Therefore, it is impossible to assess from this data what proportion of the English Wiktionary views come from language learners. Having said that, CEFR may also have some relevance for someone who is not actively learning a language (e.g., language teachers or writers who want to draft a text for a specific audience).

Another, more general problem, concerns the completeness and accuracy of CEFR labeling as such. As already mentioned, CEFR information for the highest C1 and C2 levels is rather fragmentary and likely not representative. In principle, these levels do not even have a cap on vocabulary coverage, and C2 in particular is generally thought to be an open-ended set. These reservations might also to a degree apply to level B2. This issue could be addressed in the future if more complete data on upper-level vocabulary become available. More generally, CEFR labeling as such is not necessarily consistent, fully principled, and remains necessarily arbitrary up to a point. There are also concerns about CEFR lacking specificity and nuance in its one-size-fits-all prescriptions (Weir, 2005, Foley, 2019).

The CEFR vocabulary set includes some specific and even arbitrary formulations that combine variants of multi-word expressions and lexicographic meta-words (*sth*, *sb*, etc.). The following list is just a small sample of such problematic items.

not be cut out to be sth/not be cut out for sth
be slow to do sth; be slow in doing sth
day by day/little by little/one by one, etc.
so did we/so have I/so is mine, etc.
from the 1870s/March/6.30 pm, etc. onwards

It is no surprise that such phrases find no exact matches amongst the English Wiktionary entries. Further work might explore whether CEFR entries such as these can reliably be mapped automatically to entry titles in the English Wiktionary.

Another avenue for future research might leverage the relationship between CEFR level and other known lexical variables, as discovered in studies such as the present attempt, to try to produce, adjust, or supplement CEFR-grading.

Data availability

All data generated or analysed during this study are available in the OSF repository at https://osf.io/5cj8w.

Received: 5 December 2023; Accepted: 15 February 2024;
Published online: 29 February 2024

References

Alderson JC (2007) The CEFR and the need for more research. Mod Lang J 91:659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Bergenholtz H, Johnson M (2005) Log files as a tool for improving internet dictionaries. Hermes 34:117–141

Brysbaert M, New B (2009) Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods 41:977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert M, Mandera P, McCormick SF, Keuleers E (2019) Word prevalence norms for 62,000 English lemmas. Behav Res Methods 51:467–479. https://doi.org/10.3758/s13428-018-1077-9

Cambridge University Press (2015) English Vocabulary Profile: The CEFR for English. Available from: https://www.englishprofile.org/wordlists/evp (July 26, 2023)

Capel A (2012) Completing the English vocabulary profile: C1 and C2 vocabulary. Engl Profile J 3:e1

Capel A (2015) The English vocabulary profile. Engl profile Pract 5:9–27

Çelik S (2013) Plurilingualism, Pluriculturalism, and the CEFR: Are Turkey's foreign language objectives reflected in classroom instruction? Procedia Soc Behav Sci 70:1872–1879. https://doi.org/10.1016/j.sbspro.2013.01.265

Council of Europe (2001) Common European Framework of Reference for Languages: Learning, teaching, assessment. Available from: https://rm.coe.int/1680459f97

Council of Europe (2018) Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors. Available from: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Council of Europe (2020) Common European framework of reference for languages: learning, teaching, assessment; companion volume. Council of Europe Publishing, Strasbourg, 274 pp. Available from: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Council of Europe (2023) Principles, Guidelines and the CEFR. Available from: https://www.coe.int/en/web/portfolio/principles-and-guidelines-and-cefr (November 25, 2023)

De Schryver G-M, Joffe D (2004) On how electronic dictionaries are really used. In: Williams G, Vessier S (eds.) Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004, Vol.1. Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, Lorient, 187–196

De Schryver G-M, Wolfer S, Lew R (2019) The relationship between dictionary look-up frequency and corpus frequency revisited: a log-file analysis of a decade of user interaction with a Swahili-English dictionary. GEMA Online J Lang Stud 19:1–27. https://doi.org/10.17576/gema-2019-1904-01

De Schryver G-M, Joffe D, Joffe P, Hillewaert S (2006) Do dictionary users really look up frequent words?—On the overestimation of the value of corpus-based lexicography. Lexikos 16:67–83. https://doi.org/10.4314/lex.v16i1.51504

Foley JA (2019) Issues on assessment using CEFR in the region. LEARN J Lang Educ Acquis Res Netw 12:28–48

Garlock VM, Walley AC, Metsala JL (2001) Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. J Mem Lang 45:468–492. https://doi.org/10.1006/jmla.2000.2784

Hiranburana K, Subphadoongchone P, Tangkiengsirisin S, Phoochaeoensil S, Gainey J, Thogsngsri J, Sumonsriworakun P, Somphong M, Sappapan P, Taylor P (2017) A framework of reference for English language education in Thailand (FRELE-TH)–based on the CEFR, the Thai experience. LEARN J Lang Educ Acquis Res Netw 10:90–119

Juhasz BJ (2005) Age-of-acquisition effects in word and picture identification. Psychol Bull 131:684–712. https://doi.org/10.1037/0033-2909.131.5.684

Koplenig A, Meyer P, Müller-Spitzer C (2014a) Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer C (ed.) Using online dictionaries. Lexicographica Series Maior. Walter de Gruyter, Berlin, 229–249

Koplenig A, Meyer P, Müller-Spitzer C (2014b) Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer C (ed.) Using online dictionaries. Lexicographica Series Maior. Walter de Gruyter, Berlin, 229–249

Krumm H (2007) Profiles instead of levels: the CEFR and Its (Ab)Uses in the context of migration. Mod Lang J 91:667–669. https://doi.org/10.1111/j.1540-4781.2007.00627_6.x

Kuperman V, Stadthagen-Gonzalez H, Brysbaert M (2012) Age-of-acquisition ratings for 30,000 English words. Behav Res Methods 44:978–990. https://doi.org/10.3758/s13428-012-0210-4

Lemnitzer L (2001) Das Internet als Medium für die Wörterbuchbenutzungsforschung. In: Lemberg I, Schröder B, Storrer A (eds.) Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Lexicographica Series Maior. Niemeyer, Tübingen, 247–254

Lew R, Wolfer S (2024) What lexical factors drive look-ups in the English Wiktionary? SAGE Open 14:21582440231219101. https://doi.org/10.1177/21582440231219101

Little D (2009) The European Language Portfolio: where pedagogy and assessment meet. In: 8th International Seminar on the European Language Portfolio, Graz. Bundesministerium für Unterricht, Kunst und Kultur

Longobardi E, Rossi-Arnaud C, Spataro P, Putnick DL, Bornstein MH (2015) Children's acquisition of nouns and verbs in Italian: Contrasting the roles of frequency and positional salience in maternal language. J Child Lang 42:95–121. https://doi.org/10.1017/S0305000913000597

Lorentzen H, Theilgaard L (2012) Online dictionaries—how do users find them and what do they do once they have? In: Fjeld RV, Torjusen JM (eds) Proceedings of the 15th EURALEX International Congress. Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, 654–660

Ministry of Education of the People's Republic of China (2018) China's standards of English language ability. Available from: https://cse.neea.edu.cn/html1/report/18112/9627-1.htm (January 24, 2024)

Müller-Spitzer C, Wolfer S, Koplenig A (2015) Observing online dictionary users: studies using Wiktionary log files. Int J Lexicogr 28:1–26. https://doi.org/10.1093/ijl/ecu029

Negishi M, Takada T, Tono Y (2013) A progress report on the development of the CEFR-J. In: Galaczi ED, Weir CJ (eds.) Exploring language frameworks: Proceedings of the ALTE Kraków Conference. Cambridge University Press, Cambridge, 135–163

Oxford University Press Oxford Learner's Word Lists. Available from: https://www.oxfordlearnersdictionaries.com/wordlists/ (September 22, 2023)

R Core Team (2023) R: A Language and Environment for Statistical Computing. Available from: https://www.R-project.org

Trap-Jensen L, Lorentzen H, Sørensen NH (2014) An odd couple—corpus frequency and look-up frequency: what relationship? Slovščina 2.0 2:94–113

Van der Meer G (2004) On defining: Polysemy, core meanings, and "great simplicity." In: Williams G, Vessier S (eds.) Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004, Vol.2. Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, Lorient, 807–815

Venables B (2023) codingMatrices: Alternative Factor Coding Matrices for Linear Model Formulae. Available from: https://CRAN.R-project.org/package=codingMatrices

Verlinde S, Binon J (2010) Monitoring dictionary use in the electronic age. In: Dykstra A, Schoonheim T (eds.) Proceedings of the XIV Euralex International Congress. Afûk, Ljouwert, 1144–1151. Available from: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202010/106_Euralex_2010_7_VERLINDE%20BINON_Monitoring%20Dictionary%20Use%20in%20the%20Electronic%20Age.pdf

Weir CJ (2005) Limitations of the Common European Framework for developing comparable examinations and tests. Lang Test 22:281–300. https://doi.org/10.1191/0265532205lt309oa

Weizman ZO, Snow CE (2001) Lexical input as related to children's vocabulary acquisition: effects of sophisticated exposure and support for meaning. Dev Psychol 37:265–279. https://doi.org/10.1037/0012-1649.37.2.265

Widdowson H (2015) ELF and the pragmatics of language variation. J Engl a Ling Franca 4:359–372. https://doi.org/10.1515/jelf-2015-0027

Acknowledgements

This research was funded by National Science Center, Poland, Grant Number 2020/39/B/HS2/00923.

Author contributions

Robert Lew: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, writing – original draft, writing – review & editing.Sascha Wolfer: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing – original draft, writing – review & editing.

Competing interests

The authors declare no competing interests.

Ethics approval

Ethics approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required or possible as the study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Robert Lew.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.