

John Benjamins Publishing Company



This is a contribution from ITL 175:1

© 2024. John Benjamins Publishing Company

This electronic file may not be altered in any way. The author(s) of this material is/are permitted to use this PDF file to generate printed copies to be used by way of offprints for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible only to members (students and faculty) of the author's institute. It is not permitted to post this PDF on the internet, or to share it on sites such as Mendeley, ResearchGate, Academia.edu.

Please see our rights policy at <https://benjamins.com/content/customers/rights>

For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

For further information, please contact rights@benjamins.nl or consult our website:
www.benjamins.com

Graded resources for learning and teaching foreign languages

An overview

Thomas François and Núria Gala

UCLouvain | Aix-Marseille Université

Innovative resources for teaching vocabulary have emerged during the last decades, among which the so-called ‘graded resources’, i.e. lexicons or inventories where linguistic forms have been associated with a difficulty level, having in mind a target reader. The idea of ‘grading’ in Education is not new and has evolved over time: vocabulary or text scales (grades) have been approached as an educational tool with the means available at each period of time. In this survey, our aim is to show how current approaches of graded resources for foreign language learning are rooted in the tradition of building frequency lists for education. We then synthesize the body of work that has been undertaken to design graduated resources, highlighting the different methodologies applied as well as existing resources.

Keywords: Graded linguistic resources, graded lexicons, vocabulary learning, second and foreign language teaching, CEFR

Introduction

The notion of ‘grade’ in Education, considered not as a rating but as a level, is not recent. Its roots can be found under the influence of two philosophers of the Age of Enlightenment (18th). First, John Locke stated that knowledge arises out of observation and experience and argued for children to have concrete experiences to learn. Similarly, Jean Jacques Rousseau spread his philosophy of education based on the idea of learning through concrete experience. These ideas, yet remote, led to the creation of grade schools and adapted reading material, thus putting the learner at the forefront.

In the early 19th century, the schools were what was called ‘one-room’ schools where a teacher would provide instruction to those children of various ages whose labour was not required by the family. The classrooms were frequently over-

crowded. “In the US over half of the nation’s children were receiving formal education, and more students than ever before now had access to levels of schooling previously restricted to an elite few” (de Castell & Luke, 1983:377). Textbooks were scarce and teachers struggled to devote individual attention to students. In 1847, the educator Horace Mann created the first so-called ‘graded school’ in the US, “schools designed to meet the particular learning needs of specific groups of students” (Foster et al., 2022: 9). From that on, reading standards were defined for each grade. Educators promoted the target reading levels for each class with the use of standardized reading tests.

Ideally, educational materials should be calibrated on a sample of readers representative of the population for which the texts are intended. This is the approach widely favored in the context of language assessment, in which *item response theory* has been developed to ensure the validity of the measurement of test difficulty (Woods and Baker, 1985). Since the seminal work of Guilford in 1936, numerous statistical methods have been proposed to calibrate items (see Thissen and Steinberg, 2009), the most famous of which might be the Rasch model (Eckes, 2009). However, calibrating items using this family of techniques requires a lot of time and resources, which has made it impractical for the grading of large sets of documents or vocabularies.

As an alternative to testing the materials with the target population, another approach has been to resort to expert perceptions, for instance teachers’ judgments. This consists in having experts of the measurement scale to assign a grade level to materials (e.g. texts, words, items, etc.) based on their knowledge. One of the most famous applications of this technique is the calibration process of the descriptors used in the Common European Framework of Reference for languages (CEFR). It involved 300 language teachers who sorted more than 2,000 descriptors according to their difficulty, before a Multi-Facet Rasch analysis was run to more reliably project all descriptors on a common scale. We will later discuss the limitations of such an approach, but the most obvious one is that even such “scaled teacher perceptions” (Wisniewski, 2017) might not correspond to actual learners’ development. In other words, this technique is prone to lacking empirical validity.

Relying on empirical evidence to grade linguistic materials is a third approach and a popular one today that leverages advances in fields such as learner corpus research or natural language processing. Alderson (2007) already advocated for empirical evidence to support claims of linkage of linguistic descriptors to the CEFR. More recently, the development of the English Profile project led the Cambridge research group to link words (Capel, 2010) as well as grammatical properties (O’Keeffe and Mark, 2017), based on the Cambridge Learner Corpus. However, grading words or texts based on empirical data can be traced back to

a much earlier period, in relation with the development of lexical frequency lists as well as the design of readability formulas, i.e. mathematical formulas aiming at matching readers to texts of an appropriate level. In other words, the goal of such lists and formulas was to alleviate the amount of effort required to calibrate each teaching material with a sample of readers.

In this introductory survey to graded resources for foreign language learning and teaching, our interest goes on the notion of grade as a way of considering the contents' complexity of the learning materials with regards to the skills of the learner. We discussed above three main ways to grade linguistic items — be it texts, grammatical notions, or words. Section 2 will explore more deeply the roots of the graded lexicons, mostly discussing the previous empirical attempts of assigning reliable grades to linguistic elements, and especially to words. Section 3 will then focus on recent research within the emerging field of graded lexicons, stressing how current studies are actually rooted in previous educational lexicons. The conclusion wraps this overview with some thoughts on current challenges and future research avenues for the field.

The roots of graded lexicons

The roots of graded lexicons can be traced back to the field of education and to the attempt to develop vocabulary lists. Vocabulary lists, sorted according to one criterion, enable teachers to select words to be taught as a priority and can be used to create pedagogical material. In this section, we first review the main milestones of the scholarly efforts to build frequency lists from large corpora, then we move to some of the most influential vocabulary lists for teaching a foreign language.

Principal milestones in the field of frequency lists

The most popular among criteria used to establish vocabulary lists is word frequencies estimated on a corpus. The facilitating effect of frequency on word recognition performance was first mentioned by Cattell in 1885, then confirmed empirically by Howes and Solomon (1951). Since frequent words are easier to access in the mental lexicon, either because of a lower activation threshold or a higher basic activation level (Brysbaert et al., 2000), and since it is rather easy to estimate frequencies, many existing word lists are based on this criterion. In addition, the long-established observation (Nation, 1990) that a student who knows a relatively small number of the most frequent words in the language (approximately 3,000 for English) can already understand 80 to 90% of a text has also established the dominance of frequencies in the field of lexical resources.

The origin of frequency lists is sometimes traced back to the work of L.A. Sherman (1893), a professor of English literature that was the first to use statistical analysis for the task of analyzing readability, introducing a new and objective method to choose reading materials for his students based on vocabulary. However, Thorndike's *Teacher's Word Book* (1921) was the first extensive listing of words in English (10,000 words ranked by frequency of use), built from a corpus of about 4.5 million words. It provided teachers with objective means for measuring the difficulty of words and texts, also laying the foundation for the research on the assessment of text readability. Thorndike also raised several methodological issues central to the building of frequency lists, such as the question of the validity of the estimates obtained on a specific corpus (and of the resulting classification of words) or the existence of some prevalent words in the immediate context that are poorly estimated on a corpus of written texts. In 1932, Thorndike followed up with *A Teacher's Word Book* of 20,000 Words, and in 1944 with Irving Lorge, *A Teacher's Word Book* of 30,000 words.

In contrast with Thorndike — who built his vocabulary list for school children, from varied types of texts — Rinsland (1945) established a list of 25,632 entries based on a collection of 100,000 texts written by schoolchildren. He was also the first to provide frequencies broken down by school year: for each word, the number of occurrences of that word in each grade (from 1 to 8) is given. This practice is now very popular in recent graded resources (see Section 3). Although using texts written by children to compile frequency measurements opened an interesting avenue for creating teaching materials, it does not seem to have been much pursued in subsequent work on L1 frequency resources.

Another milestone is the *American Heritage Word Frequency Book* by Carroll et al. (1971). Based on a corpus of 5.09 million words extracted from texts relative to 17 subjects and used in American schools from grade 3 to 9, this list stands out because the frequency information is provided for each of the 17 subjects and for each of the 8 levels of education. In addition to the raw frequencies, three 'original' statistical indices (Carroll, 1972), which are superior to those of previous frequency counts (Darnell and Howes, 1972), are given for each word. First, the relative entropy of words defined over the 17 subjects (D) measures the *dispersion* of frequencies. This value can be used to detect words that mostly appear in specific contexts. Next, the U value is an estimate of the 'real' frequency (frequency per million words) of the word in a theoretically infinite corpus and is normalized for dispersion. The final measure calculated is the *standard frequency index* (SFI), resulting from a logarithmic transformation of U . The database itself, however, was rapidly criticized as the corpus was not well-balanced in terms of levels and subjects.

Subsequently, the development of computational linguistics made it easier to gather larger corpora and extract frequency information from them less laboriously. The list from Kucera and Francis (1967), based on the Brown Corpus, is probably the most well-known example of this trend. The norms defined by these authors have remained one of the most popular measures of frequency among (psycho)linguists until recently (Brysbaert and New, 2009). In addition, the Brown corpus was one of the first to be automatically POS-tagged – with the TAGGIT tagger (Greene and Rubin, 1971), which was a first step towards looking beyond orthographic forms of words. A few years later, the LOB corpus was created to replicate the approach of Kucera and Francis for British English. Hofland and Johansson (1982) built a list for British English from it and compared the frequencies to those of the Brown list, introducing the idea that frequency lists can be useful for comparing different language varieties.

Among the recent frequency lists for school children or adults, we can highlight the CELEX database (Baayen et al., 1996) that was the first to recommend the use of lemma frequency: until now, the frequencies calculated have concerned forms as encountered in corpora (Brysbaert and New, 2009), possibly associated with POS-tags. Similarly to CELEX, the *British National Corpus Frequencies* (Leech et al., 2001) reports frequencies of inflected forms as well as lemmas. In addition, leveraging the presence of written and oral texts in the corpus, the authors introduced a distinctiveness index able to distinguish word profiles across language variations (e.g. oral vs. written). Intended for school children, the *Educator's Word Frequency Guide* (Zeno et al., 1995) has been established on a 17-million-word corpus of texts used in American classrooms from grade 1 to 12. It also reports a frequency distribution, for each word, over the twelve grades. Finally, the SUBTLEX project stressed the dependence of previous frequency lists on written documents and computed lists from subtitles corpora for various languages, starting with French (New et al., 2007). The authors also introduced the use of the Zipf scale as more interpretable than the logarithm of the U values from Carroll (1972).

Lists for foreign language teaching

With regards to language learners, the first efforts can be traced back to the *Basic English* by Ogden, who proposed a simplified subset of regular English words, conceived as an aid for teaching English as a second language. This approach thrived in the context of the audio-lingual method in foreign language teaching, in which the focus was put on acquiring the main phonological and grammatical structures with a basic vocabulary. For instance, the most influential 'basic' list for French in this context is *Le français élémentaire* by Gougenheim (1954).

At the same time, West (1953) published one of the most influential basic vocabularies for English: the *General Service List* (GSL) that includes 2,000 words combining judgment experts and sense frequency. Counting the frequency of the different senses of a given word (manually), thus being able to stress the most frequent one, was a great innovation to the field that has been hardly reproduced in other lists. Unfortunately, the GSL might not represent modern word usage, which has led other researchers to produce modern versions, such as the *New GSL* by Browne (2013) or the one by Brezina and Gablasova (2015).

Another ambitious project is the *English Vocabulary Profile*, which links about 7,000 entries to the 6 levels of the CEFRL (Capel, 2010). It also distinguishes between different senses of the same spelling form: the project sought to determine the order in which the various meanings are acquired. To establish the level of each word, the authors analyzed learner's data such as the Cambridge Learner Corpus (50M tokens) or the Cambridge English Corpus of First Language Use (1.2B tokens). The main weakness of this resource is the lack of frequency information, which makes it impossible to document the relative importance of words within a level.

Graded lexicons for vocabulary learning

At the beginning of the 21st century, different ideas, methods, technologies, and teaching practices converged to enable the creation of more various graded resources. In this section, we describe the various approaches that have been used to build graded lexicons, starting with a quick discussion of the Reference Level Descriptors for the CEFR (Section 3.1), then focusing on recent work that directly drawn from some of the innovations relative to frequency list building that we reviewed in the previous section (see Section 3.2). Finally, other approaches are presented in Section 3.3.

Graded resources from experts

As already discussed, behind the concept of 'graded lexicon' lies the fundamental notion of educational scale. A wide range of scales and tests have been developed in recent years to assess proficiency or achievement in languages. If we focus on foreign language teaching, one of the most significant scales is the Common European Framework (CEFR) that includes six grades of proficiency, from beginner (A1–A2), intermediate (B1–B2), to proficient (C1–C2). This scale has become highly popular across the globe, being used as a standard, even in non-European countries (Alderson, 2007). At the same time, it has not been without criticisms,

on the intuitive teacher judgements of the scale, on the relevance and validity of the level descriptors, or on the inadequacy to consider individual differences and variability among learners (Hulstijn, 2007; Fulcher, 2010). North (2020) has however undertaken to address some of these criticisms, advocating the methodological foundations of the CEFR.

A natural consequence of the popularity of the CEFR scale was the rise of interest in producing linguistic inventories graded according to it. What is probably the most ambitious initiative in this regard are the Reference Level Descriptors (RLDs) for more than 20 languages, a notable example being the EVP project (Capel, 2010). RLDs have assigned various linguistic elements (words and expressions, syntactic structures, phonemes, etc.) to CEFR levels. This classification was performed based on criteria selected by the authors for their relevance and objectivity: essentially the official descriptors from the CEFR, collective knowledge and experience of experts, teachers and examiners, and examples of learner productions deemed to be at a particular level (Beacco et al., 2008). Although popular in L2 teaching, RLDs have not been much used for other purposes,¹ especially in natural language processing, due to either accessibility reasons (paper format only, copyright), but also lack of systematicity in the format (see Pintard and François, 2020).

Graded resources inferred from educational corpora

Scholars in computational linguistics have investigated another way of grading words in reference to a proficiency scale, following a two-step procedure that considers word occurrences in graded reading materials. First, following the work of Rinsland (1945), a frequency list is built in which word distributions over the intended scale are estimated on the graded materials. Then, for each word, the distribution can be transformed into a grade level using some mathematical function that should ideally be empirically defined.

As regards to the first step, there are several examples of frequency lists that do not only provide a total frequency of use for each word, but their distribution of use across various levels. For school children, the *Educator's Word Frequency Guide* (Zeno et al., 1995) is probably the first of such a resource. Other projects rapidly flourished in different languages. For British English, we may refer to *Children's Printed WordDatabase* (Stuart et al., 2003; Masterson et al., 2010) that, in its extended version, contains 12,193 words and their distribution from Reception to Year 3 (5–9 years). For French, the Manulex resource (Lété et al., 2004)

1. A notable exception is TextInspector (<https://textinspector.com/>), which can analyze a text using the vocabulary listed in EVP.

was developed because of a need in child language studies. The authors used books from primary French schools to create a list in which the frequency of each word was estimated over three different levels: CP as an equivalent for first grade, CE1 for 2nd grade, and CE2–CM2 corresponding to grades 3rd to 5th. For instance, the verb *bourdonner* (‘to buzz’, line 2895 in Figure 1) appears in all the grades (U value estimates the frequency per million words), whereas *bourdonnant* (‘buzzing’, line 2893) appears from level CE1 onwards.

1	LEMES	NLET	SYNT	CP U	CE1 U	CE2-CM2 U	CP-CM2 U
2886	bouquet	7	NP			0,01	0,01
2887	bouquetin	9	NC	0,50		1,70	1,41
2888	bouquin	7	NC		4,03	3,12	3,17
2889	bouquiner	9	VER			0,02	0,01
2890	bourbon	7	NP			0,02	0,01
2891	bourde	6	NC			0,02	0,01
2892	bourdon	7	NC	0,61		4,12	2,94
2893	bourdonnant	11	ADJ		0,19	0,79	0,77
2894	bourdonnement	13	NC		4,13	2,46	2,91
2895	bourdonner	10	VER	3,80	13,11	5,86	8,08
2896	bourdonnot	10	NP			0,06	0,03
2897	bourg	5	NC		0,19	4,11	2,94
2898	bourgade	8	NC			0,75	0,44
2899	bourgeois	9	ADJ		0,36	1,29	1,11

Figure 1. Sample from Manulex

For Spanish, the *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años* (Martínez and García, 2004), describes the frequency of words across the six grades of primary school. Later, the LEXIN resource (Corral et al., 2009:1010) targeted very young children (in kindergarten and first grade) and estimated frequencies based on “134 reading and spelling books designed for children learning to read”. For Portuguese, Soares et al. (2014) developed a resource for 6 to 11-year-old children based on a corpus of about 3.2 million words. Interestingly, in addition to the commonly used dispersion, U frequency, and standard frequency index, they also report the contextual diversity measure, which corresponds to the number of textbooks in which a word appears divided by the total number of textbooks considered in each level. Although related to the dispersion measure, it is not totally correlated.

The same methodological approach was applied to foreign language learning, especially within the framework of the CEFRLex project. Inspired by Manulex, the project started in 2014 with the publication of the French resource FLELex (François et al., 2014). The authors collected a corpus of 28 French as a foreign language textbooks and 29 simplified readers to estimate the frequency distribution of 14,236 words across the 6 levels of the CEFR scale. Besides targeting L2 learners instead of school children as the previous psycholinguistic databases, another innovation of FLELex was to offer an alternative version of the database

including 2,037 multi-word expressions (MWE) such as *à cause de, journée-type*, etc. ('due to', 'standard day'). Later, Todirascu et al. (2019) provided a database of 4,295 verbal MWE out of which 1,098 were assigned a CEFR level.

In addition to French, other languages were targeted in the CEFRLex project. SVALex (François et al., 2016) was developed for Swedish as a foreign language and includes 15,681 words, among which 2,672 are MWE. FLELex and SVALex, as most aforementioned resources, target the receptive knowledge of L2 learners. This is the reason why the corpora used to estimate word frequency distribution are books and textbooks read by the intended audience. SweLLEx (Volodina et al., 2016) is currently the only resource in CEFRLex built with a corpus of essays to directly model productive knowledge of L2 learners. Finally, for Swedish, Holmer and Rennes (2022) introduced NyLLEx that focuses on Easy Swedish and comprises 6,668 words (68% of them overlap with SVALex whereas only 41% overlap with SweLLEx).

As a part of the CEFRLex project, similar databases were also built for English (EFLLex, Dürlich and François, 2018), Spanish (ELELex, François and De Cock, 2018), German (DAFLex, unpublished), and Dutch (NT2Lex, Tack et al., 2018). An interesting feature of NT2Lex is that two linguistic units were considered: the orthographic forms and the meanings. All the CEFRLex resources are available from the project interface² that allows for searching individual words or to analyze a whole text. For instance, Figure 2 shows how the interface allows the user to visualize and compare word distributions, for instance three adverbs that can express the idea of opposition (e.g. however) in French: *cependant* (ensure) in orange, *néanmoins* in blue, and *toutefois* in purple.

All these frequency databases offer two advantages over traditional frequency lists obtained from a very large corpus of authentic texts. On the one hand, frequencies are estimated on much smaller corpora, but on texts that are much closer to the real input that schoolchildren or L2 learners encounter during learning. Zevin and Seidenberg (2002) compared the Zeno list with the BROWN list and CELEX and noticed that Zeno norms were more predictive of word reading response latencies. They considered that such effect could be explained either by the fact that the norms were computed on children-target texts, or because they provide cumulative frequencies, which "may be a better predictor of adult performance" (Zevin and Seidenberg, 2002: 7). The experiments of Pintard and François (2020), in which they trained support vector machines to predict the CEFR level of words in the French RLD using frequency information from either FLELex or Lexique3 (New et al., 2007), also revealed that cumulative frequencies across several proficiency levels predict better expert-defined levels of words than frequency information obtained from a large subtitle corpus.

2. <https://cental.uclouvain.be/cefrlex/>

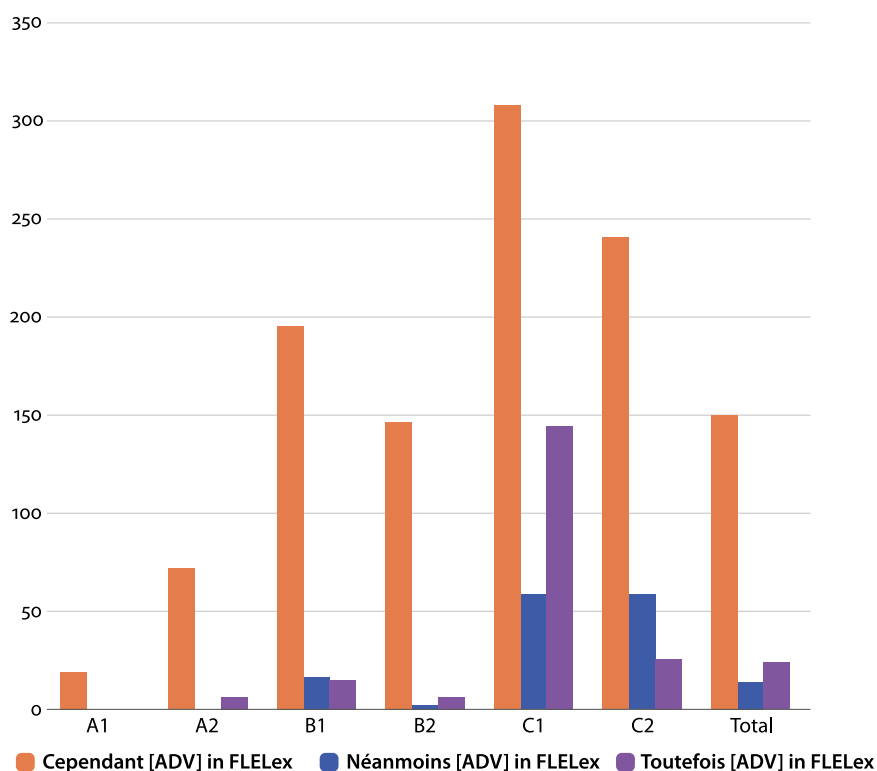


Figure 2. Screenshot from FLELex

Based on such resources, the second step to create a graded lexicon is to transform the frequency information into a single difficulty level. Several methods have been investigated to this aim. A first one simply assigns to a given lexical item the lowest level at which it was observed in the corpus (Gala *et al.*, 2014). For instance, the word *bourdonnant* ('buzzing') in Figure 1 would be assigned to the level CE1, as it is its first level with non-null probability. A slightly more nuanced technique is to consider all occurrences of a given word as a statistical series and select the level associated to a given decile or percentile (Gala *et al.*, 2014). In this trend, Alfter *et al.* (2016) employed the *significant onset of use* which consists in assigning a word to the first level whose frequency is sufficiently greater than that of the previous one. Pintard & François (2020) trained a supervised learning algorithm for transforming frequencies into CEFR levels from Beacco's RLD. We are not aware of any studies that have succeeded in validating such transformation functions in a more empirical manner, i.e. based on a large amount of learners' data. It should not be forgotten that these frequency lists are generally obtained by using corpora of different sizes, which means that a threshold could be difficult to generalize.

Other strategies to grade lexical units

The main weakness of the above approach is the need for reliably graded corpora from which to extract the word distributions over the difficulty scale. When no such corpus is available, a simpler approach consists in building a traditional frequency list over a large corpus and assigning words to instructional levels through the definition of frequency bins, assigning each bin to a grade level (e.g. Kilgariff *et al.*, 2014). This band-based method is in line with the large body of work that assesses word frequencies based on their band rank rather than their count (Crossley *et al.*, 2013), which is illustrated by the well-known web frequency analyzer Lextutor.³

Alternatively, a corpus-free way of identifying word complexity is by studying its orthographical, morphological, syntactic, and semantic properties. While it is assumed that word frequency is the most important factor in helping to identify which words are likely to be learned first (Milton, 2009), several other linguistic factors were also reported to influence word learnability (Laufer, 1997). Similarly, in her review of all independent variables that have been examined in studies interested in predicting lexical competence, Tack (2021) identified more than 300 factors related to the input, the subject, interaction between both, and the method of the study. However, only a few scholars have used word characteristics beyond frequency to build graded resources. Gala *et al.* (2014) identified 49 different linguistic features of French word forms that could help characterize complexity (e.g. word length, syllabic structure, grapheme-phoneme correspondence, morphological properties, degree of polysemy) to train a machine learning algorithm able to generalize the grading scale in Manulex and FLELex to other words. Alfter and Volodina (2018) reproduced the same approach for Swedish on 3 frequency lists, being able to generalize word grade levels based on 15 different families of features, including topic distribution, which has brought significant improvement in performance to the models. Later, Billami *et al.* (2018) also exploited word characteristics in a ranking algorithm able to rank a pair of words. They applied this algorithm to a dictionary of synonyms, getting ReSyf, a list of 57,589 French ranked synonyms according to their difficulty to be read or understood.⁴

A last approach that has been investigated is a combination of calibration with a sample of the target population and the expert annotation. It relies on crowdsourcing annotations of the difficulty of words, generally in the context of a comparison task. Thus, in Volodina *et al.* (2022), a mixed sample of 23 L2 learners and 20 experts were asked to judge micro-tasks by groups of 4 words, among which

3. The tool is available online: <https://www.lex tutor.ca/>

4. The resource is available online: <https://cental.uclouvain.be/resyf/>

participants had to select the easiest and the most difficult words. Based on their ranks, words were assigned points that were averaged over all annotators in order to get a linear difficulty score for each word. In a second step, the authors clusterized all words using 10 anchor words – whose CEFR level was already established – in order to generalize those levels to the whole annotated list of words.

Conclusion

In this synthesis, we have sought to highlight the bridges that exist between classic work on frequency lists and more recent approaches aimed at designing graduated vocabulary lists. We showed that some of the fundamental features of the current approaches, such as the use of word distributions over a difficult scale or estimating frequencies on corpora more similar to the input of target populations, is actually rooted in the long tradition of educational resources. As regards the use of L2-intended texts to estimate frequencies, this strategy is still debated in the community, in particular as it is hampered by the problem of collecting enough of such annotated data, not to mention the issue of the validity of these annotations. As a result, frequencies are less robustly estimated than in a large L1 corpora. However, a number of findings support this strategy: Brysbaert et al. (2021) showed that the words known by non-native speakers substantially differ from native speakers, whereas Pintard and François (2020) were better able to predict CECFR levels of words (assigned by experts) with frequency distributions over the CEFR, estimated on a small L2 corpus, than with L1 frequencies estimated on a large corpus.

Another characteristic of the current approach to graded resources is the desire to transform the frequency distributions obtained into a single difficulty level. Although this is undoubtedly less effective in terms of information representation, discrete values such as a level have the advantage of being more interpretable by language teaching professionals. In this regard, there is still room for improvement in the mathematical function used to transform distributions into levels. For example, one might collect training data via crowdsourcing to train deep learning models.

Finally, an interesting research perspective consists in going beyond the lexicon to grade a wider set of linguistic phenomena, such as syntactic structures, morphological properties, etc. This special issue includes several avenues in this direction. Finally, the question of how this type of graded resource can be used by L2 professionals remains open. We have seen that there are some web interfaces⁵

5. E.g. Lextutor and CEFRLex, already cited, but also <https://spraakbanken.gu.se/larka/svlp> for Swedish.

freely available, but they remain mostly used by researchers. Thus, there is still a great deal of work to be done in adapting graded resources to the realities of L2 teaching.

Funding

This work has been supported by a FNRS “Crédit de recherche” grant to the CEFRLex project (2021–2023). The research on the lexicon ReSyf was directly funded through a grant from the French ANR ALECTOR (ANR-16-CE28-0005) (2017–2021).

References

-  Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, 91(4), 659–663.
- Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E., & Pilán, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 1–7.
-  Alfter, D., & Volodina, E. (2018). Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 79–88.
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX lexical database* (cd-rom).
- Beacco, J.-C., Lepage, S., Porquier, R., and Riba, P. (2008). *Niveau A2 pour le français: Un référentiel*. Didier.
- Billami, M. B., François, T., & Gala, N. (2018). ReSyf: a French lexicon with ranked synonyms. In *Proceedings of COLING 2018*, 2570–2581.
-  Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Browne, C. (2013). The New General Service List: celebrating 60 years of vocabulary learning. *The Language Teacher*, 37(4), 13–16.
-  Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1), 65–85.
-  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. *Behavior Research Methods*, 41(4), 977–990.
-  Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supranational levels based on yes/no decision. *Second Language Research*, 37(2), 207–231.
- Carroll, J. B. (1972). A New Word Frequency Book. *Elementary English*, 49(7), 1070–1074.

- Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. Houghton Mifflin.
-  Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(1):1–11.
-  Corral, S., Ferrero, M., & Goikoetxea, E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods*, 41(4), 1009–1017.
-  Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41(4), 965–981.
-  Darnell, D. K. & Howes, D. H. (1972). Review of The American Heritage Word Frequency Book. *Research in the Teaching of English*, 6(2), 222–246.
-  de Castell, S. & Luke, A. (1983). Defining ‘Literacy’ in North American Schools: Social and Historical Conditions and Consequences, *Journal of Curriculum Studies*, 15:4, 373–389.
- Dürlich, L. & François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of LREC 2018*, 873–879.
- Eckes, T. (2009). *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Section H: Many-Facet Rasch Measurement*. In Council of Europe (Ed.).
- Foster, K. R., Graziano, L., & Squire, J. (2022). Horace Mann Charter Schools: Their Past, Present, and Promise. *Bellwether Education Partners*.
- François, T., & 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 (poster)*.
- François, T., Volodina, E., Pilán, I. & Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of LREC 2016*, 213–219.
- François, T., Gala, N., Watrin, P., & Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of LREC 2014*, 3766–3773.
- Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. *Advances in Research on Language Acquisition and Teaching. Selected Papers*, 15–26.
- Gala, N., François, T., Bernhard, D. & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. In *Proceedings of TALN 2014*, 91–102.
- Gougenheim, G. (1954). Le français élémentaire. *The French Review*, 27(3): 217–220.
- Greene, B. B. & Rubin, G. M. (1971). *Automatic Grammatical Tagging of English*. Department of Linguistics, Brown University.
- Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. Norwegian Computing Centre for the Humanities.
- Holmer, D., & Rennes, E. (2022). NyLLELex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level. In *Proceedings of LREC 2022*, 1326–1331.
-  Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of experimental psychology*, 41(6), 401.
-  Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal*, 91(4), 663–667.

-  Kilgariff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., ... & Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1), 121–163.
- Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present Day American English* (1st edition). Brown University Press.
- Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In N. Schmitt & M. McCarthy (Eds.) *Vocabulary Description, Acquisition and Pedagogy* (pp. 140–155). Cambridge University Press.
-  Lété, B., Sprenger-Charolles, L. & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, 36 (1), 156–166.
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Routledge.
- Martínez, J. A., & García, M. E. (2004). *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años*. Salamanca: Universidad Pontificia de Salamanca.
-  Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology*, 101, 221–242.
- Nation, P. (1990). *Teaching & Learning Vocabulary* (1st edition). Heinle & Heinle.
-  New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, 28(4), 661–677.
- North, B. (2020). The CEFR illustrative descriptors: Validation reference paper for researchers. [Unpublished manuscript]. Council of Europe.
- O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4), 457–489.
-  Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Short Run Press Ltd.
- Pintard, A., & François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st READI Workshop*, 85–92.
- Rinsland, H. D. (1945). *A basic vocabulary of elementary school children*. Macmillan.
- Sherman, L. A. (1893). *Analytics of Literature: A manual for the objective study of English prose and poetry*. Ginn.
-  Soares, A. P., Medeiros, J. P., Simões, A., Machado, J., Costa, A., Iriarte, A., João de Almeida, J., Pinheiro, A. P., & Comesaña, M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46, 240–253.
-  Stuart, M., Dixon, M., Masterson, J., & Gray, B. (2003). Children's early reading vocabulary: Description and word frequency lists. *British Journal of Educational Psychology*, 73, 585–598.
- Tack, A. (2021). *Mark my words! On the automated prediction of lexical difficulty for foreign language readers*. PhD University of Leuven & université catholique de Louvain-la-Neuve (Belgium).

-  Tack, A., François, T., Desmet, P. & Fairon, C. (2018). NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the BEA Workshop 2018*, 137–146.
-  Thissen, D. & Steinberg, L. (2009). Item response theory. In R.E. Millsap & A. Maydeu-Olivares (Eds.) *The Sage handbook of quantitative methods in psychology*, (pp. 148–177). Sage Publications Ltd.
- Todirascu, A., Cargill, M., & François, T. (2019). PolylexFLE: une base de données d'expressions polylexicales pour le FLE. In *Proceedings of TALN 2019*, 143–156.
- Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016). SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 76–84.
-  Volodina, E., Alfter, D., & Tiedemann, T.L. (2022). Crowdsourcing ratings for single lexical items: a core vocabulary perspective. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 10(2), 5–61.
- West, M. (1953). *A general service list of English words*. London: Longman.
-  Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning*, 67(1), 232–253.
-  Woods, A., & Baker, R. (1985). Item response theory. *Language Testing*, 2(2), 117–140.
- Zeno, S., Ivens, S.H., Millard, R.T., & Duvvuri, R. (1995). *The Educator's word frequency guide*. Touchstone Applied Science Associates.
-  Zevin, J.D., & Seidenberg, M.S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47(1), 1–29.

Address for correspondence

Thomas François
Cental, IL&C, UCLouvain
Place Montesquieu, 3, bte L2.06.04
1348 Louvain-la-Neuve
Belgique
thomas.francois@uclouvain.be
 <https://orcid.org/0000-0001-9415-7639>

Co-author information

Núria Gala
Aix Marseille Univ, CNRS, LPL
nuria.gala@univ-amu.fr
 <https://orcid.org/0000-0003-2987-0723>

Publication history

Date received: 15 June 2023

Date accepted: 30 June 2024

Published online: 13 September 2024