

INTRODUCTION

Graded resources for second and foreign language learning

David Alfter and Thomas François

University of Gothenburg | UCLouvain

The strong relationship between vocabulary knowledge and reading comprehension has been thoroughly researched and confirmed (Laufer, 1992; Hsueh-Chao & Nation, 2000). Readers should ideally know between 95% and 98% of the words in a text to adequately comprehend it (Laufer & Ravenhorst-Kalovski, 2010). Given that the larger the vocabulary, the better the comprehension of texts, foreign language curriculums should stress the need to teach new vocabulary items. As learners cannot easily learn thousands of words, it is crucial for foreign language curriculum designers, publishers of educational materials and textbooks, or even teachers to identify the most important words to teach at each stage of the learning process.

The most common answers to that challenge have been (1) to use frequency lists obtained from a large corpus of texts intended for native readers (e.g., Kucera & Francis, 1967; Leech & Rayson, 2014; Brysbaert & New, 2009) and (2) to rely on expert knowledge, such as teacher expertise or linguists' recommendations (Beacco et al., 2008; Capel, 2010). The former approach is rather straightforward to implement but suffers from well-known shortcomings. One is the observation that large corpora – in which the written mode is generally predominant – does not adequately capture the importance of some words that are prevalent in the learning environment, but rare in corpora (Gougenheim et al., 1964). In addition, while native speaker language is accepted as an ideal learning horizon, it is not the best representation of the different stages of the interlanguage of second or foreign language (L2) learners. For instance, Brysbaert et al. (2021) showed that the words known by non-native speakers substantially differ from native speakers. As regards the second approach, it offers the advantage of relating words to some educational proficiency scale, notably that of the CEFR, but its empirical foundations have been challenged (Hulstijn, 2007; Fulcher, 2010).

More recently, new trends have been investigated, such as projecting frequency lists onto the CEFR scale (Kilgariff et al., 2014) or resources in which word frequencies are estimated directly on texts intended for L2 readers and annotated

according to a proficiency scale (François et al., 2014; Dürlich & François, 2018). However, these new trends only focus on the lexicon, whereas the CEFR Reference Level Descriptors (e.g. Beacco et al., 2008; O’Keeffe & Mark, 2017) may include grammatical structures, functions, pattern of spelling development, morphemic patterns, etc.

Recent efforts have opened new avenues for research in the field of graded resources for language learning. For instance, Volodina et al. (2024) aim at developing data-driven CEFR-graded resources for grammatical or morphological patterns. In parallel, new lexical resources have used meanings as the main entry instead of the spelling form (Tack et al. 2018 ; Alfter, 2021). Acknowledging that the field of graded resources is both still in its infancy and on the verge of possible major changes, particularly as a result of the revolution of generative AI, we are convinced that it is the right time to launch a special issue devoted to this theme. This special issue is rooted in a workshop dedicated to graded resources, “GR4L2 – Building CEFR-graded resources for second and foreign language learning”, which took place at the UCLouvain in 2021 and gathered scholars from all over the world for a program of 16 presentations. Among these, five were selected for publication in this special issue as the result of a call. In addition, **Thomas François and Núria Gala** open this issue with a synthesis on current research aiming to build graded resources for learning and teaching foreign languages. This survey shows how previous efforts have been scattered across a number of disciplines, such as text readability assessment, teaching education, assessment, or natural language processing. It also emphasizes the ties between the methodology used to build frequency lists for education purposes and what we call graded resources, that keep relying on the notion of frequency.

One of the main challenges in the creation of reliable graded lexical resources is that it is a complex endeavor, particularly when it comes to assigning words to specific proficiency levels. Automating this process based on frequency or other metrics can lead to inconsistencies and inaccuracies, as word difficulty can vary significantly depending on individual learners’ backgrounds, interests, and experiences. To address this challenge, **Flor, Holtzman, Deane and Bejar**, informed by previous similar approaches (Gala et al., 2014; Alfter and Volodina, 2018), propose a two-step approach that combines manual annotation with machine learning. First, they carefully select a representative portion of the vocabulary and manually annotate it, establishing a set of reference annotations. These reference annotations serve as the ground truth for the grading task, providing a reliable basis for evaluating the accuracy of automated methods. Next, they employ machine learning algorithms to train classifiers on the manually annotated data. These classifiers are then used to automatically annotate a much larger corpus of words, expanding the scope of the graded lexical resource without sacrificing accuracy. This

approach leverages the power of machine learning while ensuring the reliability of the annotations by relying on a carefully curated set of reference annotations. **Todirascu, Cargill and François** also address this challenge, assigning grade levels to multi-word expressions (MWE) leveraging frequency information of MWE in an educational corpus. Although this approach had already been applied to single words, the authors have also evaluated different strategies with language learners.

Another aspect that is often overlooked in graded resources, and more generally in lexical lists, are collocations and multi-word expressions (MWEs) (Alonso-Ramos, 2013); yet, mastering collocations and multi-word expressions poses a significant challenge to language learners (Foster et al., 2014). These fixed or semi-fixed combinations of words often carry unique meanings or nuances that cannot be accurately inferred from the individual words alone. **Cuberos Vicente, Rosado Villegas and Mañas Navarrete** investigate the distribution of collocations in Spanish language production across different levels of the Common European Framework of Reference (CEFR). They found that simply focusing on individual collocations, or isolated features, is insufficient for predicting the overall proficiency of non-native Spanish speakers. Instead, a more comprehensive approach that considers the interplay of collocations and other linguistic factors is required. In line with this finding, **Todirascu, Cargill and François** present a new graded resource for French learners that meticulously compiles and categorizes MWEs, providing valuable support for acquiring the nuances and idiomatic expressions of the language. This resource aims to assist learners in developing a deeper understanding of French lexico-grammar and usage, ultimately enhancing their overall language proficiency.

While lexical knowledge is an essential component of language proficiency, most graded resources tend to emphasize the lexical level, neglecting the structural and grammatical aspects of language. This narrow focus can limit learners' ability to grasp the nuances of grammar and structure, which are crucial for producing fluent and natural language. **Uchida, Arase and Kajiware** recognize this gap and present a dataset that goes beyond lexical annotation, providing CEFR-annotated sentences. Thus, their research takes place in an area that is sometimes referred to as sentence readability (Pilan et al., 2014), which has been clearly insufficiently researched. They provide what might be the first freely CEFR-annotated dataset for English, CEFR-SP, comprising 7,511 carefully selected sentences. In a second step, this dataset is used to investigate and better illustrate the characteristics of sentences in each CEFR level.

Finally, words are not isolated constructs; knowledge of one word may facilitate learning and understanding of related words (Bauer and Nation, 1993; Schmitt and Zimmerman, 2002). **Volodina, Mohammed and Lindström Tiede-**

mann introduce the Swedish Word Family, a graded resource for learners of Swedish as a second language. This resource is organized around word families, which are groups of words that share a common root. They describe the construction of the resource and its applicability for L2 research through several case studies. This resource provides a valuable tool for language researchers studying Swedish vocabulary acquisition and usage. It can be used to identify patterns of word relationships, track vocabulary growth over time, and analyze the impact of different pedagogical approaches.

To conclude, this special issue on graded resources for language learning marks a significant turning point in the field, showcasing cutting-edge research that addresses long-standing challenges and opens new avenues for innovation. The collected works represent a paradigm shift in how we approach language learning materials, moving beyond simplistic frequency-based lists to more sophisticated, multi-dimensional resources. By tackling issues such as the reliable grading of lexical items, the integration of multi-word expressions and collocations, the development of comprehensive CEFR-annotated datasets, and the organization of vocabulary around word families, these studies lay a robust foundation for the next generation of graded language learning resources.

The research presented here not only offers innovative solutions to existing problems but also pushes the boundaries of what's possible in language resource development. The combination of manual expertise with machine learning techniques, the focus on learner-centric data, and the holistic approach to language proficiency demonstrate a mature and nuanced understanding of the complexities of language acquisition. Moreover, this body of work underscores the critical importance of interdisciplinary collaboration, bringing together insights from corpus linguistics, natural language processing, educational psychology, and second language acquisition theory.

Looking forward, the field of graded resources is likely to continue evolving, particularly with the advent of generative AI. Future research may focus on creating even more personalized learning resources, integrating various linguistic aspects, and exploring how these tools can be effectively implemented in both traditional and digital learning environments. As the field progresses, it will be crucial to continue bridging the gap between theoretical linguistics, educational practice, and technological innovation to create increasingly effective language learning solutions.

As a final word, we would like to thank our reviewing committee for their excellent work in reviewing the articles for this special issue. The list of members of the committee can be found at the end of this issue.

Funding

Open Access publication of this article was funded through a Transformative Agreement with University of Gothenburg.

References

-  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.
- Alfter, D. (2021). Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective (The dissertation is published here: <https://gupea.ub.gu.se/handle/2077/66861>). University of Gothenburg.
- Alonso-Ramos, M. (2013). Colocaciones, diccionario y corpus de aprendices. In *Eugenio Coseriu, in memoriam: XIV Jornadas de Lingüística* (pp. 57–72). Servicio de Publicaciones.
-  Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.
- Beacco, J.-C., Lepage, S., Porquier, R., & Riba, P. (2008). *Niveau A2 pour le français: Un référentiel*. Éditions Didier.
-  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.
-  Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(1):1–11.
- 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.
-  Foster, P., Bolibaug, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, 36(1), 101–132.
- 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., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'Élaboration de français fondamental, 1er degré. Étude sur l'établissement d'un vocabulaire et d'une grammaire de base. Nouvelle édition, refondue et augmentée*.

- Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–30.
-  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.
-  Kilgarrieff, 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*. Brown University Press.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15.
-  Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Palgrave Macmillan.
-  Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.
- 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.
-  Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the Ninth workshop on innovative use of NLP for building educational applications*, 174–184.
-  Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL quarterly*, 36(2), 145–171.
-  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.
-  Volodina, E., Alfter, D., & Tiedemann, T. L. (2024). Profiles for Swedish as a second language: Lexis, grammar, morphology. In *Huminfra Conference*, 10–19.

Address for correspondence

David Alfter
GRIDH, University of Gothenburg
david.alfter@gu.se
 <https://orcid.org/0000-0002-1232-6118>

Co-author information

Thomas François
Cental, IL&C, UCLouvain
thomas.francois@uclouvain.be
 <https://orcid.org/0000-0001-9415-7639>

Publication history

Date received: 15 June 2023

Date accepted: 30 June 2024