

ABSTRACT

Towards a graded lexical inventory in L2 Spanish: insights from productive vocabulary knowledge

Rocio Cuberos Vicente¹ and Barbara De Cock²

¹Universitat de Barcelona; ²Université catholique de Louvain
rociocv89@gmail.com ; barbara.decock@uclouvain.be

It is widely acknowledged that vocabulary is a major component of native and non-native linguistic competence (Hulstijn, 2015). Since its publication, the CEFR has been widely used in developing language syllabuses and proficiency tests and has, thus, become accepted as the standard for grading L2 proficiency (Figueras, 2012). However, when it comes to setting vocabulary goals, concerns about how to grade lexical repertoire still arise today (Tarrés, 2017; François, Volodina, Pilán & Tack, 2016). For Spanish, the reference level descriptors of the Plan Curricular del Instituto Cervantes (Instituto Cervantes, 2006), based on the CEFR, provide lexical lists per level, but the number of lemmas is limited. This lack of systematization leads L2 teachers to the challenge of identifying the most important words to teach - and thus to assess - at each stage of the learning process.

Most research seeking to address this issue has examined word frequencies in L1 corpora (e.g., Sebastián, Cuetos, Carreiras & Martí, 2000), frequency distribution of receptive vocabulary knowledge across different CEFR levels (e.g., Dürlich & François, 2018) and expert knowledge (e.g., García Salido, 2017; Instituto Cervantes, 2006). This study, by contrast, seeks to explore the pedagogical potential of analysing frequency distribution of productive vocabulary knowledge across different CEFR levels.

The data comes from the GRERLI-ES2 corpus (Tolchinsky, Aparici, Rosado, & Cuberos, 2023) which includes 280 texts, comprising in total 34,212 tokens, produced by 70 non-native speakers of Spanish (L1 = Arabic, Korean, Chinese). All participants produced four texts. After watching a silent video about conflict situations at school, they were asked to tell orally and in writing a personal experience about a similar situation in which they had been involved (narrative text), and to discuss orally and in writing the kind of problems that were shown in the video (expository text) (v. Berman & Verhoeven, 2002). Participants were distributed into four levels of competence (A2, B1, B2 and C1) which was assessed by two expert judges with an extensive experience in assessing competence with the CEFR. Texts were tagged using the part-of-speech tagger FreeLing. As regard the lexical frequency estimation, the normalized frequency per million words was obtained.

In this presentation, we examine the overall distribution of tokens and lemmas within our dataset across proficiency levels. Also, we report hapax legomena and the number of lemmas that were observed more than 10 and 5 times. Examples of the most frequently used lemmas at different CEFR levels are provided. Finally, the frequency list is compared to the data available in the L2 Spanish graded lexical resource ELELex (François & De Cock, 2018), which includes the word frequencies observed in textbook reading activities across CEFR levels. This comparison allows us to contrast receptive and productive vocabulary knowledge at different levels of language proficiency.

References

- Berman, R. A., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text production abilities: Speech and writing. *Written Language and Literacy*, 5(2), 1-172.
- Dürlich, L. & François, T. (2018). A graded lexical resource for learners of English as a foreign language. In *LREC 18, Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 873-879.
- Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477-485.
<https://doi.org/10.1093/elt/ccs037>
- François, T. & De Cock, B. (18th May, 2018). ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. Communication. *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*. Louvain-la-Neuve.
- François, T. Volodina, E., Pilán, I., & Tack, A. (2016). SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC 16, Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 213-219.
- García-Salido, M. (2017). La frecuencia de corpus como criterio para nivelar colocaciones léxicas. *Études romanes de Brno*, 38(2), 29-49. <https://doi.org/10.5817/ERB2017-2-3>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers. Theory and research*. John Benjamins.
- Instituto Cervantes. (2006). *Plan curricular del Instituto Cervantes. Niveles de referencia para el español*. Instituto Cervantes.
- Sebastián-Gallés, N., Martí, M. A., Cuetos, F., & Carreiras, M. F. (2000). *LEXESP: Léxico informatizado del español*. Edicions de la Universitat de Barcelona.
- Tarrés, I. (2017). Aplicaciones de enfoques léxicos a la enseñanza comunicativa. In F. Herrera, *Enseñar léxico en el aula de español: el poder de las palabras* (pp. 35-46). Difusión.
- Tolchinsky, L., Rosado, E., Aparici, M., & Cuberos, R. (2023). *Spanish GRERLI-ES2 Corpus (data file)*. <https://doi.org/10.21415/2DJF-JH34>