**The effect of lexical frequency on word duration: analyzing corpus data in Spanish and English**

The present study investigates the effect of lexical frequency on word duration in Spanish and English. Previous studies have found that word duration shortens in English in more frequent words (e.g., Gahl, S. 2008; Gahl 2009; Lohmann 2018). The shortening of frequent forms may correspond to articulatory routinization (Bybee 2001; Newmeyer 2006), but evidence showing that in homophone pairs (e.g., *thyme – time*), the item with higher frequency is shorter than the infrequent one reveals that frequency effects on duration may not be due to repetition of a phonological form but to lemma frequency effects instead (Gahl, S. 2008; Gahl 2009).

The present study aims to replicate the frequency effect found in duration in English (Gahl, S. 2008; Gahl 2009; Lohmann 2018) and explore the effect of lexical frequency on word duration in Spanish. This study analyzes corpus data, with the English data coming from the *Free ST American English Corpus* and Spanish data from the *Crowdsourced high-quality Argentinian Spanish Speech Data Set* (Guevara-Rukoz et al. 2020). The English data were 2806 recordings of cellphone conversations, and the Spanish data consisted of 1928 recordings of random Spanish sentences recorded by Argentinian speakers. Lexical frequency was measured using movie subtitle frequencies (see New et al. 2007). The data was analyzed using two Bayesian linear regressions with duration as the outcome variable and speech rate and orthographic length as fixed predictors.

The results replicated the frequency effects previously found in English and expanded them to Spanish. English frequent words were found to be shorter than infrequent ones when orthographic length and speech rate were controlled for (Figure 1). We found similar results in Spanish, where the results exhibited a negative linear relationship between lexical frequency and word duration (Figure 1). The findings have implications for L2 acquisition of Spanish, since assumptions that Spanish is syllable-timed would not predict differences in duration based on lexical frequency.

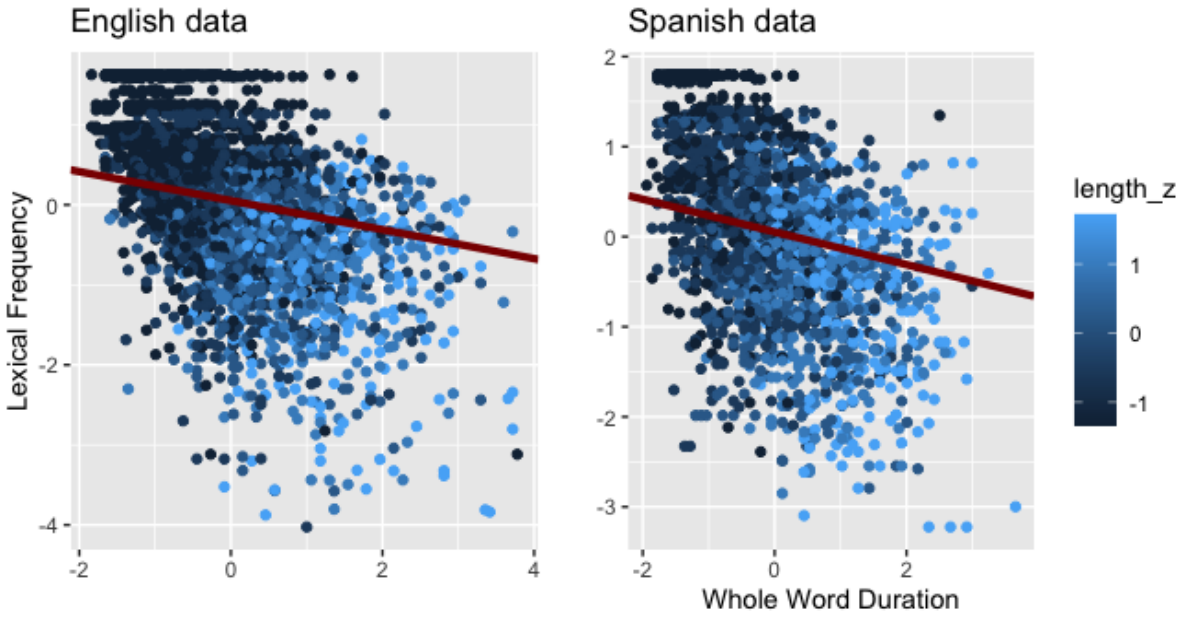


Figure 1. Whole word duration in English (Left Panel) and Spanish (Right Panel) as a function of lexical frequency and orthographic length (length_z) with the most plausible line of best fit.

## References


Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511612886.

Gahl, S. 2008. "*Time and Thyme* Are Not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech." *Language*, 84 (3): 474–96. https://doi.org/10.1353/lan.0.0035.

Gahl, S. 2009. "Homophone Duration in Spontaneous Speech: A Mixed-Effects Model." *Undefined*. https://www.semanticscholar.org/paper/Homophone-Duration-in-Spontaneous-Speech.

Guevara-Rukoz, Adriana, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. "Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech." In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 6504–13. Marseille, France: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.lrec-1.801.

Lohmann, Arne. 2018. "Cut (n) and Cut (v) Are Not Homophones: Lemma Frequency Affects the Duration of Noun–Verb Conversion Pairs." *Journal of Linguistics*, 54 (4): 753–77. https://doi.org/10.1017/S0022226717000378.

New, Boris, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. "The Use of Film Subtitles to Estimate Word Frequencies." *Applied Psycholinguistics*, 28 (4): 661–77. https://doi.org/10.1017/S014271640707035X.

Newmeyer, Frederick J. 2006. "On Gahl and Garnsey on Grammar and Usage." *Language*, 82 (2): 399–404. https://doi.org/10.1353/lan.2006.0100.