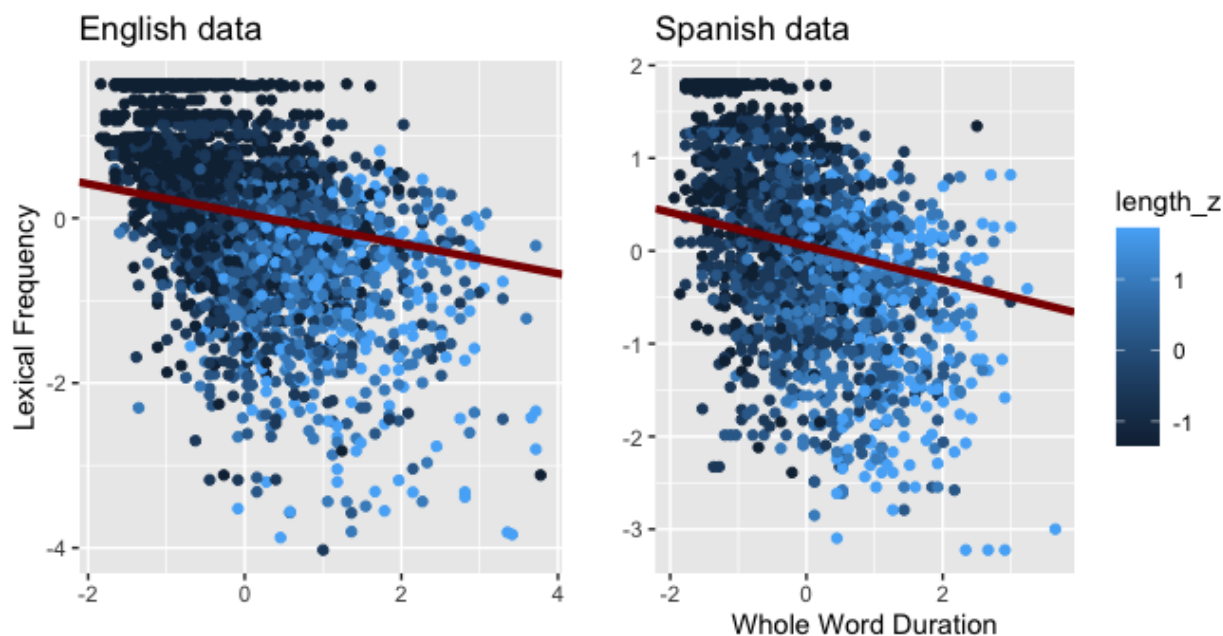


Evidence of a frequency effect in Spanish L1 word duration

In linguistic research, the nature of the relationship between linguistic sound representation and extra-linguistic factors such as lexical frequency have been a point of debate, and whether these factors impact sound representation on a language-specific or universal basis. For instance, previous studies have found evidence that more frequent words in English tend to have shorter duration. These studies have involved homophone pairs in English (e.g., *time* versus *thyme*) and have found that the lower frequency member of the pair (thyme) typically has a longer duration than the high frequency word (time). Previous research has argued that this result rules out articulatory routinization effects as a possible explanation for shorter duration of frequent words, and suggests that lemma frequency is a more fitting explanation (Gahl 2008, 2009; Lohmann 2018). However, it is unclear whether the frequency effect documented in English is language specific or could be observed cross-linguistically.

To address this question, we investigate whether there is evidence of similar shortening of more frequent words in Spanish using methods in corpus phonetics. This study analyzes two corpora, 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 corpus consists of 2806 recordings of cellphone conversations, while the Spanish data was made up of 1928 recordings of random Spanish sentences recorded by Argentinian speakers. Lexical frequency was measured using movie subtitle frequencies (see New et al. 2007). We ran two Bayesian multilevel regression models in which duration was the outcome variable and speech rate and orthographic length were population level predictors.

The results point to a replication frequency effects previously found in English, and expanded them to Spanish, in which the parameter estimates of the models suggest that the effect size is similar in both languages. Specifically, 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 a negative distribution of parameter estimate from the Bayesian model is taken as evidence that increasing frequency is associated with shorter whole word duration (Figure 1). These findings have implications for theories of universal sound representation, since the present data supports the notion that lexical frequency similarly effects are similar cross-linguistically.



\begin{figure}
\caption{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.} \end{figure}

References

- 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>.
- . 2009. “Homophone Duration in Spontaneous Speech: A Mixed-Effects Model.” *Undefined*. <https://www.semanticscholar.org/paper/Homophone-Duration-in-Spontaneous-Speech%3A-A-Model-Gahl/f3639c21294abb0f01fe2e0b792aa4b30ddaf89f>.
- Guevara-Rukoz, Adriana, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheak-mungkol 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>.