Prosody drives alternations: Evidence from a 61 Million Word Corpus of Brazilian Portuguese Film Subtitles

Kevin Tang[1], Andrew Nevins[1], Michael Becker[2]

[1]University College London & [2]Indiana University Bloomington, Department of Linguistics; [1]kevin.tang.10@ucl.ac.uk, [1]a.nevins@ucl.ac.uk, [2]michael.becker@phonologist.org

Introduction

Token frequency is widely claimed to be a conditioning factor in alternations in the lexicon [Bybee, 1995, 2003, Huback, 2007, Coetzee and Kawahara, 2010]: lexical items that are used more often are more likely to undergo phonological processes. In this paper we examine the plural morphology of Brazilian Portuguese, and show that the correlation between token frequency and alternations are epiphenomenal, and in fact depend on prosodic shape.

Brazilian Portuguese nouns ending in -Vw, e.g. [ˈsaw] "salt", [ʒuɾ.ˈnaw] "newspaper", [ˈmɛw] "honey", [ˈsɛw] "sky", [ʃɐ.ˈpɛw] "hat", often, but not always, show an alternation in the plural, whereby the final glide becomes palatal, e.g. [ˈsajʃ], [ʒuɾ.ˈnajʃ], [ˈmɛwʃ], [ˈsɛwʃ] and [ʃɐ.ˈpɛjʃ] / [ʃɐ.ˈpɛwʃ]. In Becker et al. [2012] it is proposed that the conditioning factor determining whether a noun will participate in such alternations is prosodic: monosyllables are preferentially protected, a trend confirmed in large-scale nonce word tasks. In the present study, we set out to test whether token frequency would also be a predictor of alternation rates for existing words in the lexicon, given the claim in Huback [2007] that frequent words alternate more often in Brazilian Portuguese.

Method

Problem

- No frequency corpora of Brazilian Portuguese of a reasonable size are available
- SUBTLEX film subtitle frequencies are excellent predictors of behavioral task measures for English [Brysbaert and New, 2009], French [New et al., 2007] and Dutch [Keuleers et al., 2010]
- Motivated us to do the same for Brazilian Portuguese

Datamining

- Opensubtitles.org (over 1,900,000 subtitles as of Dec 2012)
- Circumvented the download limit per IP address per day by changing IP addresses through public proxy servers
- Downloaded 25,303 zip files with films/TV series subtitles

Post-processing

Preliminary selection and cleaning

- Each downloaded zip files contains one to two subtitle files (in different formats, e.g. .srt, .sub, and in multiple discs) and an information file
- Used only .srt files, since they are the majority and could avoid potential duplicates from the different formats
- Left with 26,627 .srt files
- Removed irrelevant information in the subtitle files such as the subtitle number per line and its start time and end time

Removing duplicates

- Duplicates could skew the overall frequency.
- Duplicates could come in different forms due to 1) different translators, 2) different released formats: a single release (1-disc) or in multiple parts (multiple-discs)
- The most important part of post-processing
- Two techniques were employed:
  1. Kullback-Leibler divergence (D_{KL}) [Kullback and Leibler, 1951], which is a non-symmetric measure of the difference between two probability distributions.
  2. K-means clustering algorithm [MacQueen et al., 1967] - a simple unsupervised learning clustering algorithms
- Identify duplicates:
  1. D_{KL} was calculated on the distributions of each file against those of the rest of the files. Duplicates will have very low D_{KL}
  2. The cut-off D_{KL} value for being duplicates was manually calibrated to ensure that files were not misidentified as duplicates
- Select unique files from duplicates:
  1. Applied the K-means clustering algorithm to sort the duplicates into two clusters based on total word frequency (high and low)
  2. The high frequency group was chosen to capture the 1-disc file over the multiple-disc files, and if only the multiple-disc files were available, it would maximize the size of our corpus by picking the largest disc.
  3. Then, the files with fewest words with a frequency of one were chosen, as they were likely to contain fewer typos
  4. 12,353 unique files were identified.

Removing Non-Portuguese files

- Errors are often made by the uploaders by uploading the wrong translation or including the original subtitles.
- To filter out these non-Brazilian Portuguese files, we employed a language detection model [Shuyo, 2010] with an above-99% precision for 53 languages
- The model calculates language probabilities from features of spelling using a naïve Bayesian model with character n-gram as well as language profiles generated from Wikipedia abstract xml.
- The model filtered out 249 files, the remaining 12,104 files have the probability of at least 99.7% of being in Portuguese

Compiling and filtering

The 12,104 files were compiled to produce a corpus with a list of words, their frequency and contextual diversity (CD) (which is a measure of the number of subtitle files that a word has occurred in). A few filters were used to further clean our corpus.

- Filter out some web URLs and e-mail addresses.
- Filter out words that do not consist only of Brazilian Portuguese graphemes ’áâãàéêíóôõúçüabcdefghijklmnopqrstuvwxyz’
- Filter out words with CD of 2 and below

Yielded a corpus with 61,609,241 tokens and 136,147 types

Acknowledgements

We would like to thank **Marc Brysbaert** & **Emmanuel Keuleers** for their support and advice on creating this corpus.

Corpus URL

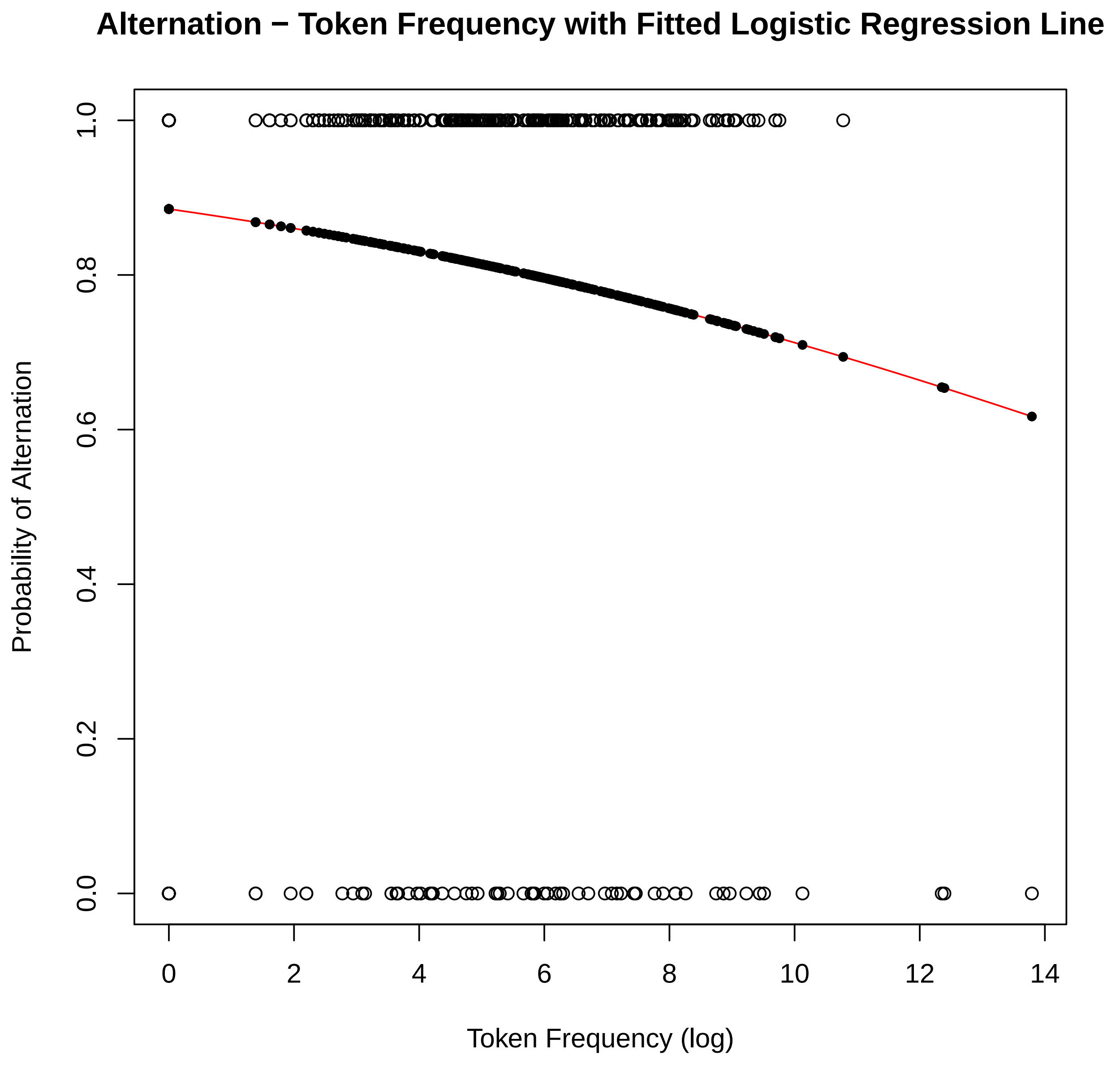
Different versions of the corpus (with different filters) with an interactive interface are available at `http://crr.ugent.be/subtlex-pt-br/`

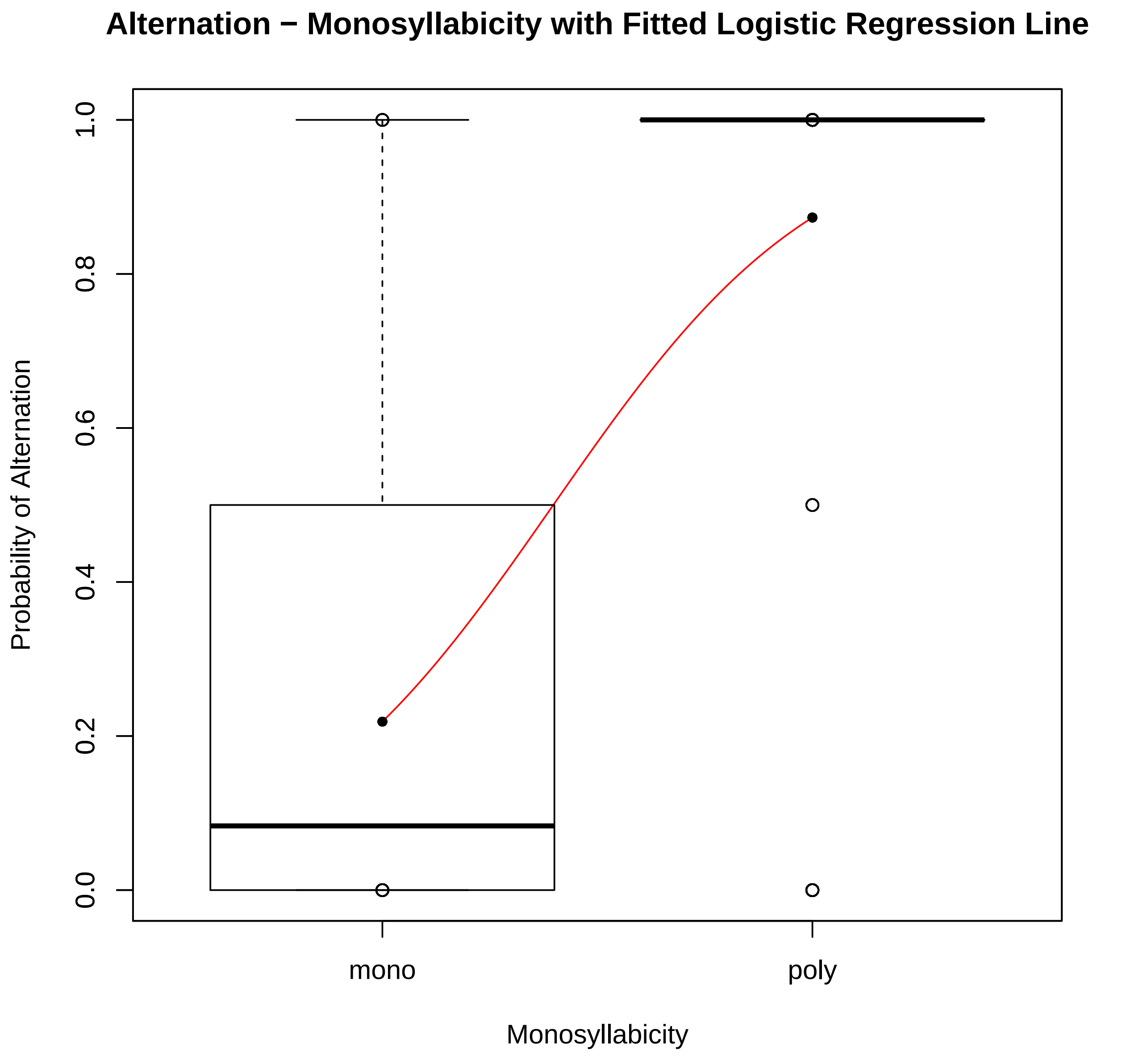
Analyses

We performed a statistical analysis using *glm()* in *R* and model comparisons using ANOVA between a superset model and subset models.

387 existing w-final words in the lexicon and their alternation rates taken from Becker et al. [2012] were tested against our corpus, of which the frequency of 313 words were found. Many models were tested with variations in including and excluding the zero frequency items in the data set, and the three predictors: shape(monosyllabicity), vowel laxness and token frequency.

Three models were created: a superset with three predictors, shape(monosyllabicity), laxness and token frequency; and two subsets: excluding a) token frequency and b) monosyllabicity. For all these w-final words, model comparison revealed that while shape, a two-level variable encoding monosyllables, and polysyllables made a significant improvement, $\chi^2(1) = 55.709$, $p < .0001$, token frequency makes no significant improvement, $\chi^2(1) = 0.15367$, $p > .1$. Further modeling showed that while token frequency was significant on its own, $p < .05$, it became irrelevant once monosyllables were taken out, $p > .1$. Together they suggest that token frequency makes no significant improvement in prediction above and beyond the simple binary shape variable.





Conclusion

This study adds to the growing body of work suggesting that frequency-tracking alone is unlikely to condition learners’ generalizations about the patterns governing morphophonological alternations, perhaps because learners have an implicit knowledge that usage frequencies may come and go with the wind, while prosodic shapes are more stable.

Furthermore, our creation of a very large subtitle corpus for Brazilian Portuguese, openly available and in a standardized format, will remain accessible as a potentially valuable resource for a number of researchers in adjacent fields.

References

M. Becker, L.E. Clemens, and A. Nevins. A richer model is not always more accurate: the case of French and Portuguese plurals. *Lingbuzz*, 2012. URL `http://ling.auf.net/lingBuzz/001336`.

M. Brysbaert and B. New. 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, 2009.

J. Bybee. Regular morphology and the lexicon. *Language and cognitive processes*, 10(5): 425–455, 1995.

J. Bybee. *Phonology and language use*, volume 94. Cambridge University Press, 2003.

A.W. Coetzee and S. Kawahara. Frequency and other biases in phonological variation. *Ms., Michigan University and Rutgers University (submitted for publication in Natural Language and Linguistic Theory)*, 2010.

A. Huback. *Efeitos de freqüência nas representações mentais*. PhD thesis, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

E. Keuleers, M. Brysbaert, and B. New. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, 42(3):643–650, 2010.

S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

B. New, M. Brysbaert, J. Veronis, and C. Pallier. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661, 2007.

Nakatani Shuyo. Language detection library for Java, 2010. URL `http://code.google.com/p/language-detection/`.