A Phoneme-weighted Levenshtein Distance: A Measure of Cross-linguistic Phonological Similarity

Abstract

Many studies in bilingualism have explored how form similarity between translation equivalents (i.e. cognateness) impacts language processing. Cognates range from being identical to sharing only a few key features. Although there exists some objective measures of orthographic similarity between word-forms, there is still no representative equivalent for phonological similarity. In infant bilingual research, a measure that reflects perceptual similarity of spoken words as presented in infant-directed speech would be of great value. A widely used metric for word similarity is Levenshtein distance [@levenshtein1966binary]. This measure calculates word-form similarity by computing the smallest total number of substitutions, insertions and deletions needed to change one word to the other. However, a challenge when defining phonological similarity is that some phonemes (e.g. [k] and [g]) are perceptually more similar than others (e.g. [k] and [o]). The standard Levenshtein distance calculation treats each change equally. Thus it often overestimates phonological distance between words. We are developing an adaptation of Levenshtein distance for phonological representations, where edit operations are weighted according to the degree of phonemic feature changes involved, type of change (e.g. vowel or consonant insertion) and position in the word. To estimate the weights, we are collecting behavioural data in an auditory translation elicitation task. Monolingual adult participants with no exposure to the target language will be asked to guess and produce the translations of auditorily-presented words. The only way for participants to guess the correct translation is by mapping the presented unknown phonological form onto the phonology of its known translational equivalent. Cross-linguistic phonological similarity will be operationalised as the probability that participants correctly guess the corresponding translation. The higher the probability, the higher the perceptual similarity. Our weighted similarity metric will provide a useful tool for identifying phonological similarity between words in infant-directed speech.

Introduction

Methods

Computing cognateness: Phonological distance across translation equivalents

Behavioural task: Transaltion elicitation task

[@levenshtein1966]

Results

Discussion

Appendix

Appendix 1: Session info

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.2
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] papaja_0.1.0.9942 ggplot2_3.3.0     citr_0.3.2        stringr_1.4.0
## [5] tidyr_1.0.2       magrittr_1.5      dplyr_0.8.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.4.6     pillar_1.4.3     compiler_3.6.3   later_1.0.0
##  [5] tools_3.6.3      digest_0.6.25    evaluate_0.14    lifecycle_0.2.0
##  [9] tibble_3.0.1     gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.5
## [13] shiny_1.4.0.2    yaml_2.2.1       xfun_0.13        fastmap_1.0.1
## [17] withr_2.1.2      knitr_1.28       vctrs_0.2.4      grid_3.6.3
## [21] tidyselect_1.0.0 glue_1.4.0       R6_2.4.1         rmarkdown_2.1
## [25] purrr_0.3.4      scales_1.1.0     promises_1.1.0   ellipsis_0.3.0
## [29] htmltools_0.4.0  assertthat_0.2.1 mime_0.9         xtable_1.8-4
## [33] colorspace_1.4-1 httpuv_1.5.2     stringi_1.4.6    miniUI_0.1.1.1
## [37] munsell_0.5.0    crayon_1.3.4
```