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

First Author[1] & Second Author[1,2]

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

Many studies in bilingualism have explored how form similarity between translation equivalents (i.e. cognate- ness) 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 (**???**). 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 Lev- enshtein 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 col- lecting 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 trans- lation. 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.

Keywords: keywords

Word count: X

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

#Methods

Translation Elicitation task

Participants. Data collection took place from 04th June, 2020 to 28th June, 2020. We collected data from 105 participants (M_{age} = 21.58, SD_{age} = 3.09, $Range_{age}$ = 2-33). 72 participants were British English native speakers living in United Kingdom (0 female), and 72 participants were Spanish native speakers living in Spain (0 female). Participants in UK were recruited via Prolific (5£ compensation) and SONA (compensation in academic credits). Participants in Spain were contacted via announcements in Faculties, and were compensated 5€ or an Amazon voucher for the same value. Participants were asked to complete the experiment in a quiet place with good internet connection. We exluded data from participants that a) self-rated their oral and/or written skills in a second or third language as higher than 4 in a 5-point scale (n = 1), b) were diagnosed with a language (n = 2)[1], or c) did not contribute more than 80% of valid trials (n = 1).

Procedure. Participants accessed the study from a link provided by Prolific or SONA and completed the experiment from a browser (Chrome or Mozilla). First participants were informed about the aims of the study and gave informed consent for participating. Second, participants answered a series of questions about their demographic status, their language background, and the set up they were using for completing the

[1] We originally planned to exclude participants that reported any visual impairment that glasses would not correct This item was phrased as "Do you have normal or corrected-to-normal VISION? (Yes/No)" in English, and as "¿Tienes problemas de VISIÓN que unas gafas o lentes de contacto NO corrijan? (Sí/No)". However, the proportion of Spanish participants that reported visual impairment was unplausibly large (n = 6, 18.18%). This is possibly due to these participants using glasses dayly and not having read the item until the end, where it is indicated that the use of glasses is considered as normal vision.

study. Third, participants completed the experimental task. BEfore the task, participants were informed that they would listen to a series of pre-recorded words in Catalan or Spanish (English participants) or only Catalan (Spanish participants). They were instructed to listen to each word, guess its meaning in English (English participants) or Spanish (Spanish participants), and type their answer as soon as possible. English participants were randomly assigned to the list of Catalan or Spanish trials. Spanish participants completed the list of Catalan trials.

Design. Each trial started with a yellow fixation dot presented during one second on the center of the screen over a black background. After one second, the audio started playing while the dot remained being displayed until the audio offset. Upon the offset of the fixation point and audio, participants were prompted to write their answer by a ">" symbol. Typed letters were displayed in the screen in real time to provide visual feed-back to participants. Participants were allowed to correct their answer. Then, participants pressed the RETURN key to start and new trial. Participants contributed a total of 9536 trials (5901 in Catalan, 3635 in Spanish). The task took approximately 15 minutes to be completed.

Stimuli.

Data analysis

Results

Discussion

78

References