GENERATING A BILINGUAL LEXICAL CORPUS USING INTERLANGUAGE NORMALIZED LEVENSHTEIN DISTANCES

Amanda Post da Silveira[1]
Jan-Willem van Leussen[2]

[1]Donders Centre for Cognition (Radboud University Nijmegen)
[2]Amsterdam Centre for Language and Communication (University of Amsterdam)
psyphon.ap@gmail.com, jwvanleussen@gmail.com

ABSTRACT

Finding large numbers of target items for phonetic and phonological experiments can be a time-consuming and error-prone task. Using freely available tools and data, we have generated a bilingual corpus with the specific aim of investigating the processing and perception of stress in second-language (L2) words. Normalized Levenshtein distances between orthographic and phonemic transcriptions of Brazilian Portuguese (BP) and American English (AmE) translation word pairs were used to automatically generate similar and dissimilar word pairs. Frequency data from corpora were used as a metric of familiarity. To test if these generated metrics correspond to speakers' representations, BP L1 speakers of AmE L2 rated the word pairs on orthographic and phonological similarity, and indicated their familiarity with the English words. Results showed a high correlation between subjective ratings and the computed similarity and familiarity values of the bilingual corpus. We conclude that automatically constructed bilingual corpora such as ours, combined with simple string similarity metrics, are a valid and useful tool for experimental research into L2 (word stress).

Key-words: Phonetics, Corpus Linguistics, Psycholinguistics, normalized Levenshtein distances, L2 segmental categorization, L2 word stress

1. INTRODUCTION

Many factors may influence the speakers' representations of L2 word stress: similarity to L1 words, familiarity of the L2 words, matches and mismatches in stress regularities between the two languages, and more [1]. One method to investigate the relative weight of these factors is to systematically vary them on target items in a behavioral experiment. However, manually inventing large numbers of valid target items can be a very laborious and error-prone work. This paper describes an effort to automate this laborious task. We have constructed a bilingual Brazilian Portuguese-American English corpus with the aim of providing phonemically controlled target items for experiments testing BP speakers' L2 representation of English word stress. All target items were equisyllabic *word pairs* consisting of an English word and its Portuguese translation. The experimental design demanded these translation pairs to be divided among three binary conditions: i) disyllabic versus trisyllabic, ii) matching versus mismatching stress patterns, and iii) similar versus dissimilar in terms of segmental string, where segmentally similar words were cognates and segmentally dissimilar were non-cognates. Table 1 gives example pairs over these three factors.

Table 1: Examples of word pairs distributed over the three factors (stress match, number of syllables, cognate status)

	Matching stress	Mismatching stress
Cognate, 2 syll.	*pollen* (AmE) *pólen* (BP)	*traitor*(AmE) *traidor* (BP)
Cognate, 3 syll.	*container*(AmE) *contêiner* (BP)	*animal* (AmE) *animal* (BP)
Non-cognate, 2 syll.	*splinter* (AmE) *farpa* (BP)	*rifle* (AmE) *fuzil* (BP)
Non-cognate, 3 syll	*castaway* (AmE) *náufrago* (BP)	*cucumber* (AmE) *pepino* (BP)

The aim of our corpus was to facilitate finding large numbers of pairs for each condition.

2. METHOD AND TOOLS

2.1 Corpus construction

The first step was to obtain two lists per language, one with disyllabic and one with trisyllabic words. These were downloaded, along with their frequency per million, from two corpora: the ASPA corpus [2] for Brazilian Portuguese, and CELEX [3] for English. Both corpora allow searching for words with a specific number of syllables. The English word list already contained transcriptions in X-SAMPA (Extended Speech Assessment Methods Phonetic Alphabet) format [4]. For the Portuguese words, we used the BP grapheme-to-phoneme module of eSpeak (www.espeak.org) to generate SAMPA transcriptions. Automatic transcriptions were sample checked and corrected by the authors.

Next, we located equisyllabic semantic pairings in these two word lists. We mined English-Portuguese translations in a dump of the English Wiktionary [5]. If a word and its Portuguese translation were both present in our lists of English and Portuguese words, respectively, this word pair was added to our output list of semantically close pairs.

Having obtained word pairs with orthographic and phonemic transcriptions, we used the *normalized Levenshtein distance* metric [6] to calculate orthographic and phonemic distances between the members of each pair. The Levenshtein distance [7] between two strings is defined as the minimum number of *edits* required to change one string into the other, where each edit is one of *deletion*, *insertion* or *substitution* of a character. This can be calculated equivalently for orthographic and phonemic strings. As an example, the word pair AmE *minister* [mɪnɪstəɹ] ~ BP *ministro* [miˈnistɾu] has an orthographic distance of two, and a phonemic distance of five (the transcriptions are in SAMPA format):

```
Ld(minister, ministro) = 2
minister➔ministr      1 (delete 'e')
ministr➔ministro      2 (insert 'o')
Ld(mInIst@R, minist4u) = 5
mInIst@R➔minIst@R    1 (substitute /ɪ/ for /i/)
minIst@R➔minist@R    2 (substitute /ɪ/ for /i/)
minist@R➔minist@4    3 (substitute -/ɹ/ for /ɾ/)
minist@4➔minist4      4 (deletion of /ə/)
minist4➔minist4u      5 (insertion of /u/)
```

For any pair of strings, the length of the longest string is an upper bound on the Levenshtein distance (substituting all characters and then deleting any superfluous characters from the result). Pairs of longer words will therefore generally be more distant. To correct for this bias, the *normalized* Levenshtein distance [6] divides the number of edits by the length of the longest string, so that all distances are a number between zero and one. Normalized Levenshtein distance (nLd) is then defined as follows:

nLd (w1,w2) = 1 -levenshteinDistanced(w1,w2)
/ max(length(w1),length(w2))

Under this definition, the word pair *minister* [mɪnɪstəɹ] ~ *ministro* [miˈnistɾu] has an *orthographic* nLd of (*1 - 2/8* =) 0.75 and a *phonemic* nLd of (*1 - 5/8* =) 0.375. Following [7], we adapted the orthographic similarity to account for the use of diacritics, which are much more common in BP than in AmE. Insertions or deletions of diacritics on a grapheme (or equivalently, substituting a grapheme for a counterpart with a diacritic) were assigned an edit cost of only 0.5, as in [8]. For the orthographic pair *replica* (AmE) ~ *réplica* (BP) this yields an orthographic similarity value of (1-(0.5/7 ≈) 0.9286.

We likewise adapted *phonemic* similarity: in phonemic pairs such as [mɪnɪstəɹ] ~ [miˈnistɾu], we felt that phonological similarity as judged by L2 speakers may be underestimated by the normalized Levenshtein metric. The three substitutions (twice /ɪ/ for /i/ and once /ɹ/ for /ɾ/) concern sounds that are phonetically similar, and are not distinguished in perception by many BP speakers of L2 English [9, 10]. To improve our measure of phonemic similarity, we pre-processed the AmE transcriptions to more closely match their categorization by BP speakers. English phonemes lacking from the BP inventory were replaced by their closest BP counterparts: e.g., the short lax close front vowel /ɪ/ was replaced with the long (tense) close front /i/. By using pre-categorized L2 phonemes according to L1 ones, the new interlanguage Levenshtein values for the same word pair *minister* [mɪnɪstəɹ] ~ *ministro* [miˈnistɾu] are: *orthographic* inLd of (*1 – (2/8)* = 0.75) and a *phonemic* inLd of (*1 – (2/8)* = 0. 75). This interlanguage Levenshtein distance using pre-categorized segments according to L1-specific L2 perception is, to our knowledge, original to this study.

With these definitions in place, we were able to assign phonemic and orthographic similarity values to each word pair in our corpus, in the range of 0 (maximally distant/different) to 1 (identical). The resulting list of word pairs contained 5,758 pairs. These could then be easily sorted and searched by: i) relative frequency of each member of the pair, ii) stress pattern, iii) word length (number of syllables, iv) number of letters, v) number of phonemes), vi) orthographic and vii) phonemic similarity.

2.2 Participants

In order to investigate whether these automatically generated metrics of similarity and familiarity corresponded to the representations of Brazilian speakers of English, we ran a lexical judgment experiment. Participants were 18 native speakers of Brazilian Portuguese (10 females and 8 males, ages between 21-62) who speak English as a second language, and were staying in the Netherlands for academic purposes or vacation. They were either paid for their participation or volunteered to participate without compensation. Their experience with English varied from 10 to 40 years and was based on instruction in regular school and in private language courses in Brazil.

Participants performed a vocabulary test previously to the actual experiment, the X_Lex205 [11], a lexical decision task in which participants have to decide whether words and nonsense words are existing words in English. The scores range from 0 to 5,000. Informants were included in the experiment only if they achieved a score of at least 3,500 words. Their average score in the X_Lex205 test was of 3,915 words.

2.3 Stimulus list composition

For the experiment, 104 cognate and non-cognate English-Brazilian Portuguese word pairs were selected. Orthographic similarities were defined in terms of interlanguage normalized Levenshtein similarity (inLd) for orthography and phonology, as described previously in this paper. A range between 1 to 0.7 nLd defined English cognate word candidates (similar in form and meaning to a counterpart Brazilian Portuguese word) and a range of 0.1 to 0.6 defined English non-cognate word candidates based on the average calculation of orthographic and phonological nLd. The lexical frequency range used was a low one, from 0 to 10 occurrences per million. A 2 x 2 x 2 design was created by the systematic combination of the variables “number of syllables” (2-syllable and 3-syllable words), “cognate status” (cognates or non-cognates words) and “stress match in L1 and L2” (words with stress matching or mismatching in L1 and L2), defining 8 word categories. In total, the stimulus list consisted of 104 stimuli: 16 words per cognate category and 10 words per non-cognate category.

2.4 Task

The same 104 word pairs were presented in print to all 18 participants. They answered to the following questions: 1) How similar in orthography are the following word-pairs to each other? 2) How close in pronunciation are the following word-pairs to each other? and 3) How familiar are each of the following English words to you?

Participants were asked to judge the English-Brazilian Portuguese translation pairs based on a Likert Scale [12] from 1 to 7: 1, being the least similar or familiar; and 7, being the most similar and familiar cases.

3. RESULTS

The total of 1,872 subjective ratings per question were obtained and a total number of 5,616 answers for the whole lexical judgment experiment. Cronbach’s alpha was calculated for the average responses to the three questions. Results showed a high correlation, $\alpha = 0.888$ *(N=312),* indicaing that the mean of responses of each word pair was similar across the three questions. For example, if a word pair *cucumber* (AmE) – *pepino* (BP) had a mean response of 1.5 for orthographic similarity, the mean for phonological similarity tended to be close to 1.5. Likewise, the judgment on the English word, *cucumber,* tended to be low, around 1.5, which means that the word was overall evaluated by subjects as unknown or less familiar.

The interclass correlation values calculated the inter-participant agreement for questions 1 (orthographic similarity), question 2 (phonological similarity), and question 3 (familiarity). The results of the statistic tests were highly significant for Question 1 *(r = 0.985, p <0.001, F(103,1854) = 83.775);* Question 2 *(r = 0.981, p < 0.001 F(103,1854) = 75.469);* Question 3 *(r = 0.853 p < 0.001, F(103,1751) = 10.658).* The high interclass correlations found here indicate that participants gave similar answers to the same items across all questions.

The subjective ratings were later correlated with corpora based values obtained via interlanguage normalized Levenshtein distances for orthographic and phonological similarities, and with log word frequencies based on CELEX for English. *Pearson correlation values* were highly significant for Questions 1 (orthographic similarity between L1-L2 word pairs), *r = 0.961, p<0.001*; Question 2 (phonological similarity between L1-L2 word pairs), *r = 0.812, p<0.001*; and Question 3 (familiarity with English words), *r=0.354, p<0.001.*

4. DISCUSSION

Results showed that the judgments of the word pairs were very consistent across participants. Participants provided almost the same judgments on lexical similarity to the same word pairs, reflecting that they perceived clear patterns of orthographic and phonological similarity between the pairs they judged. We account such low variability in subjective lexical judgment to the fact that participants belonged to the same L2 proficiency group – advanced learners – according to the vocabulary test scores we used to select them for this experiment.

The results of this investigation indicate that the use of the interlanguage normalized Levenshtein distances matched almost perfectly with the subjective ratings, which means that the distance values achieved with this method corresponded closely to L2 speakers' orthographic and phonological representations of the L2 words they judged.

The average correlation values for familiarity with the words were also high, although lower than the judgments on similarity. The correlation of subjective judgments on lexical familiarity and corpus frequency correlated considerably poorly. Presumably, because the subjective experience with the L2 lexicon does not match closely to the frequency in which words appear in the lexicon overall, especially in which concerns low frequency words, such as the ones used in this lexical judgment task.

5. CONCLUSIONS

We have illustrated a method to automatically generate target items for Phonetic and Neuro-Psycholinguistic experiments involving orthographic or phonological similarity between words. By linking various free sources and tools (two freely accessible corpora and an open-source multilingual dictionary; open-source grapheme-to-phoneme conversion software) we created a large word pair corpus over which phonological and orthographic similarity could be calculated using string edit distance metrics. Additionally, we introduced a novel method for calculating phonemic similarity as perceived by L2 speakers. These similarity metrics over pairs correlate very well with similarity judgments elicited in a behavioral experiment for these same pairs.

We conclude that our word pair corpus provides an effective tool for L2 experiment design. More generally, we have illustrated how valuable linguistic datasets and novel methods may be derived by linking existing tools and databases.

ACKNOWLEDGMENTS

We would like to thank Prof. Dr. Vincent van Heuven, Prof. Dr. Niels O. Schiller and Dr. Johanneke Caspers for suggestions to this corpus and on statistical analysis. The first author wishes to thank Erasmus Mundus and CNPq for funding her PhD project.

REFERENCES

[1] Post da Silveira, A.; van Heuven, V., Caspers, J., Schiller, N.O. (2014). Dual activation of word stress from orthography: The effect of the cognate status of words on the production of L2 stress. *Dutch Journal of Applied Linguistic*, 3 , 2, 170–196.

[2] Cristófaro-Silva, T., de Almeida, L.S., Fraga, T. (2005). ASPA: A Formulação de um Banco de Dados de Referência da Estrutura Sonora do Português Contemporâneo. *Proceedings XXV Congress of Brazilian Society of Computing Science*, São Leopoldo, RS, Brazil.

[3] Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (1995). The CELEX Lexical Database. Release 2 [CD-ROM]. *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia.

[4] Wells, J.C., (1997). 'SAMPA computer readable phonetic alphabet'. In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B. SAMPA available on website (www.phon.ucl.ac.uk/home/sampa)

[5] Wiktionary (en.wiktionary.org) . Date of the dump: 20/08/2012.

[6] Schepens, J., Dijkstra, T. and Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15, 1, 157 166.

[7] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals, *Doklady Akademii Nauk SSSR*, 163(4):845-848, 1965 (Russian). English translation in Soviet Physics Doklady, 10(8):707-710, 1966.

[8] Heeringa, W. (2004). Measuring dialect pronunciation differences using Levenshtein distance. [Doctoral Thesis], University of Groningen.

[9] Bion R. A. H.; Escudero, P.; Rauber, A. S.; Baptista, B. O. (2006). Category formation and the role of spectral quality in the perception and production of English front vowels. *Proceedings of INTERSPEECH*, Pittsburgh, 1363-1366.

[10] Nobre-Oliveira, D. (2007). The effects of training on the learning of American English vowels by native Brazilian Portuguese speakers. [Doctoral Thesis] Universidade Federal de Santa Catarina.

[11] Meara, P., Milton, J.(2006). X-Lex: the Swansea Vocabulary Levels Test. In Coombe, C., Davidson, P. and Lloyd D. (eds) Proceedings of the 7th and 8th Current Trends in English Language testing (CTELT) Conference, vol 4. UAE; TESOL Arabia, pp 29-39.

[12] Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1–55.