Modelling Phonetic and Phonological Variation with ‘Small’ Data: Evidence from Kaqchikel Mayan
Ryan Bennett & Kevin Tang (Yale University), kevin.tang@yale.edu

Motivation: The age of ‘big data’ has facilitated large-scale research on the effects of lexical statistics on phonological and phonetic patterning. However, it remains true that the majority of the world's languages are under-resourced, being spoken by relatively small populations. This renders large-scale lexical research extremely difficult, if not impossible, for most lesser-studied languages (Anand et al. 2011). Using the under-resourced language Kaqchikel (Mayan) as a test case, we provide a proof-of-concept that psycholinguistic and phonetic norms computed from less-than-ideal corpora are sufficient for predicting fine-grained differences in phonetics and phonology, such as the robustness of phonemic contrasts, allophonic variability, and word-frequency effects on sub-phonemic detail.

Methods: There are no currently existing structured corpora of spoken or written Kaqchikel, apart from dictionaries (e.g. Patal Majzul 2007). We therefore compiled two new corpora, one written and one spoken, for our study. The text corpus was constructed from existing religious texts, spoken transcripts, government documents and educational books. The resulting corpus contains 1 million word tokens. This is far from ideal. First, at least 16mil words are needed to reliably estimate the corpus frequency of low-frequency words (Brysbaert and New 2009). Second, while the genres the corpus covers are reasonably balanced, they are not as speech-like as spoken transcripts. The spoken corpus consists of roughly 4 hours of spontaneous spoken Kaqchikel, collected in Guatemala in 2013 by one of the authors. The recordings are free narratives, told as monologues by 16 native speakers of Kaqchikel (19-84 years old, mean=33, median=28; 6 male). The entirety of the corpus was transcribed by a native-speaker linguist, and included in the computation of lexical statistics below. The corpus is not large: it includes only 40,000 or so word tokens. A subset of the corpus (~1 hour) was submitted to forced alignment (Gorman et al. 2011) for the purposes of acoustic analysis.

Analyses: Three analyses were conducted over these two corpora:

Phonemic Contrasts: Kaqchikel dialects differ as to how many tense-lax contrasts are preserved in the vowel system (maximally /a ~ ä/, /i ~ ü/, /o ~ õ/, /u ~ ū/ and /e ~ ë/; umlauts denote lax vowels; Patal Majzul et al. 2000). If a system preserves any number of tense-lax contrasts, it will preserve the /a ~ ä/ contrast. All systems with more than one lax vowel have an /i ~ ü/ contrast. Systems with more than two lax vowels all have both /o ~ õ/ and /u ~ ū/. The dialect typology suggests the following hierarchy of resistance to merger: /a ~ ä/ > /i ~ ü/ > /o ~ õ/ > /u ~ ū/ > /e ~ ë/ (where “>” means “more resistant to merger”).

Given that written Kaqchikel preserves all five tense-lax contrasts, we were able to compute the Functional Load (FL; Hockett, 1955) of those contrasts using our text corpus. FL is a measure of how much a language makes use of a particular contrast. It is widely believed that contrasts with low functional loads are more likely to be lost, since such contrasts serve to distinguish relatively few words. Crucially, this relationship is more than simply correlative (see Wedel et al. 2013 for an extensive study modelling mergers in eight different languages). We found that FL closely correlates with resistance to vowel merger among tense-lax pairs across Kaqchikel dialects: /a ~ ä/: 4.51 > /i ~ ü/: 2.10 > /o ~ õ/: 0.90 > /u ~ ū/: 0.45 > /e ~ ë/: 0.21 (numbers are FL values on the scale of 10^{-3}). Tense-lax contrasts with higher FL tend to be preserved across dialects.

Allophonic distributions: Lax /ä/ shows a wide range of phonetic variation in Kaqchikel. Rather than being simply a centralized variant of its lax counterpart /a/, as would be expected from other tense-lax pairs, lax /ä/ is realized as any of [a æ ɨ], depending on the dialect. We explored the distribution of vowel qualities in our spoken corpus using visual inspection of density plots (contour maps) and simple k-means clustering over a Lobanov-normalized F1 x F2 space. We observed that most lax vowels are realized as slightly centralized variants of their tense counterparts. However, lax /ä/ has three distinct acoustic realizations, corresponding roughly to [u i ui]. Of these variants, [u] and [i/ui] appear to be in free variation. But why is lax /ä/ so acoustically variable, while other lax vowels are relatively stable? We speculate that allophonic raising of lax /ä/ from [u] to [i/ui] is motivated by contrast preservation pressures: the /ä ~ a/ contrast has the highest functional load of any tense-lax pair in Kaqchikel, and so allophonic raising of lax /ä/ serves to enhance the acoustic distinctiveness of
one of the most important vowel contrasts in the system.

**Word frequency effect:** We conducted a corpus-based analysis of durational reduction in spontaneous Kaqchikel speech. It has long been observed that words are often reduced in production when they are contextually predictable. This reduction takes a range of forms, including word/syllable shortening, segment/syllable deletion, and others (Hooper 1976; Whalen 1991; Aylett and Turk 2006). Focusing on the overall predictability of a word, more frequent words are more predictable than less frequent words, therefore they should also be shorter in duration. Token frequency was extracted from the text corpus. Word duration was extracted from the spoken corpus as our dependent measure in the mixed model (using `lmer()` in the `lme4` library in R) (11432 words; 2195 sentences; 16 speakers). All numeric variables were log10 and z-transformed to reduce collinearity and improve normality. To control for the effect of word length, we included the number of segments and the number of syllables in the word as predictors; these were further transformed into two principle components to reduce collinearity. To control for the effect of phrase-final lengthening, we included the position of the word in the sentence. To allow for speaker- and sentence-level variation, we included speaker and sentence ID as random effects. We followed the modelling strategy recommended in Barr et al. (2013): we began with a saturated model, with fully crossed and fully specified random effects. We used a data-driven approach to determine the random effect and fixed effect structure of our model, guided by which step of inclusion or exclusion of a predictor would lead to the best next model. The resultant model has the maximal effect structure justified by model comparison, as shown below. Despite the extensive controls, Token Frequency remained a highly significant predictor of word duration ($B = -0.104$, $t = -7.01$, $p < 0.001$), correlating negatively with duration as expected. It is the third most important predictor after Word Length (PC1, PC2) ($B = 0.47, 0.48$).

$$\text{Duration} \sim \text{Token Frequency} \times \text{Word Length (PC1)} \times \text{Word Length (PC2)} \times \text{Word Position in Sentence} + (\text{Token Frequency} + \text{Word Length (PC1)} + \text{Word Length (PC2)} + \text{Word Position in Sentence | Speaker ID}) + (\text{Token Frequency} + \text{Word Length (PC1)} + \text{Word Length (PC2)} + \text{Word Position in Sentence | Sentence ID})$$

**Conclusion:** In summary, we established a proof-of-concept for under-resourced languages: it is possible to model the effect of lexical factors on phonetic and phonological variation with less-than-ideal spoken and text corpora. Specifically, our analyses of Kaqchikel highlighted that (a) FL can account for cross-dialect variation in the preservation of tense-lax vowel contrasts as well as the target of allophony and its distribution; (b) the classic word frequency effect on duration remains robust even in a sparse corpus, and in a highly-inflecting, morphologically-complex language like Kaqchikel.


