Appendix A: Processing of written corpus

Tokenization

The corpus was pre-processed in a series of steps in order to remove 'noise' (punctuation, non-alphabetic characters, etc.). First, email addresses and website links were removed using regular expressions. Second, we replaced graphemes with a space if they were not included in the set of graphemes used in either the Kaqchikel or Spanish orthographies. We retained Spanish graphemes because Spanish words are sometimes used in Kaqchikel written texts, although historically Spanish words are sometimes written in the Kaqchikel orthography as well (e.g., *kwenta* < Spanish *cuenta* 'account'). The Kaqchikel graphemes are *aeiouäëïöübchjklmnpqrstzwxy'*. The graphemes that are used by Spanish but not Kaqchikel are *áéíóúdgñ*. Thirdly, the texts were tokenized into words using spaces as separators, yielding a large word list.

Grapheme-to-phoneme conversion

Grapheme-to-phoneme conversion was used to translate orthographic words of Kaqchikel into a phonemic representation. Both Kaqchikel and Spanish have 'shallow' orthographies, with close correspondence between graphemes and phonemes. The phonemic transparency of these orthographies made it possible to create a rule-based grapheme-to-phoneme conversion script in Python to carry out this task (as opposed to a probabilistic converter trained on transcribed words). We classified each word in the cleaned corpus into one of four categories, depending on the graphemes it contained: a) Kaqchikel, a word that could only be Kaqchikel; b) Spanish, a word that could only be Spanish (containing uniquely Spanish graphemes or grapheme sequences like *g*, *rr*, *ce*, etc.); c) Either, a word that could be either Kaqchikel or Spanish, because it contains only graphemes that are shared by both languages; and d) Mixed, a word that contains both exclusively Kaqchikel graphemes and exclusively Spanish graphemes. The Either words are treated as belonging to the Kaqchikel category for practical purposes, based on the assumption that most word types in the corpus are in Kaqchikel rather than Spanish. After a word had been classified, we applied grapheme-to-phoneme conversion, using different conversion rules for the Kaqchikel and Spanish words.

Word filters

The written corpus includes various word forms which contain errors (typos, OCR errors, other digitization errors, etc.). To filter out word forms of this type, as well as words-forms which are not clearly words of Kaqchikel, we applied a number of filters to the word list, eliminating:

1. Words which were classified as being Mixed or Spanish.
2. Words containing no vowels.
3. Words consisting of only one vowel, with the exception of /e/ (3PL.ABS) and /i/ (the Spanish conjunction *y* 'and,' which is used frequently in Kaqchikel; Brown, Maxwell, & Little, 2010, p. 197).
4. Words consisting of multiple vowels and no consonants.

Adaptation of phonemic inventory

The phonemic inventory of Patzicía Kaqchikel includes only 6 vowels, tense /a e i o u/ and lax /ə∼ɨ/ (orthographic *ä*, Majzul, Matzar, & Serech, 2000, p. 35). In this respect

Patzicía Kaqchikel—the focus of our perception study—differs from Sololá Kaqchikel and other varieties of the language which have retained a larger number of tense∼lax vowel contrasts (e.g., Bennett, 2016, to appear and references there). The standard Kaqchikel orthography represents all 5 lax vowels *ä ë ï ö ü* explicitly: Since this orthography over-represents the number of phonemic contrasts actually present in Patzicía Kaqchikel, the phonemic transcription of word-forms in the corpus was converted to a representation which merged all tense∼lax vowel contrasts with the exception of *a ä*. About 800 new homophonic word pairs were created as a result of this vowel merger. Manual inspection suggested that most of the merged word pairs were not actually distinct lexical items to begin with, but rather alternative ways of writing the same words across dialects which may have different vowel systems. For this reason, merged homophonic word pairs were considered to be a single lexical item when calculating lexical frequency.

References

Bennett, R. (2016). Mayan phonology. *Language and Linguistics Compass*, *10*(10), 469–514. doi: 10.1111/lnc3.12148

Bennett, R. (to appear). La tensión vocálica en el kaqchikel de Sololá, Guatemala: un estudio preliminar. Mexico City: Colégio de México.

Brown, R. M., Maxwell, J., & Little, W. (2010). *La ütz awäch?: Introduction to Kaqchikel Maya language*. Austin, TX: University of Texas Press.

Majzul, L. F. P., Matzar, P. O. G., & Serech, C. E. (2000). *Rujunamaxik ri Kaqchikel chi': variación dialectal en Kaqchikel*. Ciudad de Guatemala, Guatemala: Cholsamaj.