Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan)

Ryan Bennett, Kevin Tang, and Juan Ajsivinac Sian

To appear in *Laboratory Phonology*

Abstract

This paper investigates the relationship between speech perception and linguistic experience in Kaqchikel, a Mayan language of Guatemala. Our empirical focus is the perception of stop consonants (plain, ejective, and implosive) in this language. Drawing on an AX discrimination task, a corpus of spontaneous spoken Kaqchikel, and a novel text corpus, we make two claims. First, we argue that speech perception is mediated by phonemic representations which include rich acoustic detail drawn from prior phonetic experience, as in Exemplar Theory. Second, segment-level distributional patterns also condition speech perception: the perceptual distinctiveness of a given pair of phonemes is affected by their functional load as well as their overall contextual predictability. These factors have an effect on discrimination even for relatively fast response times, suggesting that top-down effects of linguistic experience may occur quite early in the timecourse of speech processing. We take this result to indicate that distributional factors like functional load may affect speech perception by shaping low-level perceptual tuning during linguistic development.

This study replicates and extends some key findings in speech perception in the context of a language (Kaqchikel) which is both structurally and sociolinguistically different from the majority languages (like English) which have served as the basis of most work in the speech perception literature. At the practical level, our research provides an illustration of some methods for conducting corpus-based laboratory phonology with lesser-studied and under-resourced languages.

1 Speech perception and linguistic experience

The perceptual similarity of any two speech sounds depends, to a great extent, on their raw acoustic similarity. However, speech perception is also mediated by the native language(s) of the hearer. The functional organization of speech sounds within a language's phonological system strongly determines whether a given pair of segments will be well-discriminated by speakers of that language (Trubetzkoy 1939; see Best 1995, Best et al. 2001, Sebastián-Gallés 2005, Boomershine et al. 2008 for references and recent discussion). For example, [d ð] are present in both English and in Spanish. In English these sounds are contrastive, as attested by minimal pairs like /be͡id/ 'bade' vs. /be͡ið/ 'bathe'. In Spanish [ð] is instead a conditioned allophone of the phoneme /d/, e.g. [deðo] 'finger' vs. [ese ðeðo] 'that finger' (e.g. Harris 1969). This difference in the function of [d ð] has consequences for speech perception:

English speakers, who rely on the [d ð] contrast to distinguish word meanings, are better at discriminating these sounds than Spanish speakers, for whom [ð] is simply a predictable variant of /d/ (Boomershine et al. 2008; see too Harnsberger 2000, 2001). These and related findings demonstrate that prior linguistic experience plays a significant role in conditioning the perception of speech sounds.

Even fine-grained details of linguistic experience, based on statistical properties of a hearer's native language, may substantially impact speech perception (see Cutler 2012 for an overview). Phoneme discrimination, for instance, appears to be sensitive to the specific acoustic parameters associated with each phoneme category in the hearer's language (e.g. Kuhl & Iverson 1995 and section 5.2). More recent research has suggested that the statistical structure of the lexicon may also influence native language speech perception (e.g. Kataoka & Johnson 2007, Yao 2011: Ch.2, Hall et al. 2014, Hall & Hume submitted, Vitevitch & Luce 2016 and references there; see Hall et al. submitted for related discussion). Among other factors, segment-level measures such as phoneme frequency and functional load (discussed in section 5.3) seem to contribute to the relative discriminability of different phoneme pairs. The precise mechanism behind such effects is not well-understood at present, a point we return to in section 7.

It thus seems clear that statistical properties of a hearer's native language may influence speech perception. However, we believe that the full generality of these findings has not yet been established, particularly with respect to lexical effects on speech perception. A large proportion of speech perception studies—perhaps most such studies—involve experiments with listeners who are native speakers of majority languages like English, French, Dutch, Japanese, and so on (e.g. Cutler 2012: 4). This sampling bias might be unremarkable, if not for the fact that the languages most commonly used in speech perception research also share a number of structural and sociolinguistic properties. To give one example, the European languages most often used in speech perception research typically belong to the Germanic or Romance branches of the Indo-European family. The morphological structure of these languages is characteristically analytic or fusional rather than agglutinating. Since perfect minimal pairs should, intuitively, be less common in languages which tend toward longer and/or more complex words, it remains unclear whether statistical measures which refer to minimal pairs (e.g. functional load) should have the same importance in languages with relatively agglutinative morphology (see also Wedel et al. 2013, Hall et al. submitted).[1] Similar questions arise with respect to neighborhood density and word-frequency effects in agglutinating languages, as longer words are likely to have fewer lexical neighbors and (possibly) low overall corpus frequencies (e.g. Zipf 1935, Yao 2011: Ch.2, Vitevitch & Luce 2016).

At the sociolinguistic level, a preponderance of work in speech perception has been conducted with listeners who are highly educated and literate in their native language. Apart from general concerns about whether results obtained with such populations are really generalizable (e.g. Henrich et al. 2010), the bias in speech perception studies toward literate speakers of Indo-European languages is potentially relevant for understanding how phoneme-

[1] A few studies have examined functional load in languages with fairly agglutinative morphology, such as Japanese, Korean and Swahili (Oh et al. 2013, 2015). To our knowledge, no existing studies have addressed the relationship between functional load and speech perception in languages of this morphological type.

level lexical statistics interact with speech perception. It has sometimes been suggested that lexical items lack a phoneme-level encoding altogether, being stored instead with strictly gestural and/or syllabic encoding (e.g. Browman & Goldstein 1986, 1989, 1992, etc.; Port & Leary 2005, Silverman 2006, 2012, Lodge 2009, Ladefoged & Disner 2012, Tilsen 2016; cf. Dunbar & Idsardi 2010, Hyman 2015). To the extent that such models of lexical storage can account for phoneme-level statistical effects in speech perception, they would presumably attribute such effects to phonemic awareness, itself an artifact of literacy in an alphabetic writing system. Against this backdrop, further studies of speech perception among populations with non-alphabetic writing systems, or simply low literacy rates, are clearly needed. It is not our place here to adjudicate between these views, only to highlight the fact that answering such questions will require a more diverse sample of speakers and languages than currently exists in the speech perception literature.

In this article we explore how statistical measures derived from the lexicon (such as pairwise functional load) affect stop consonant discrimination in Kaqchikel, a Guatemalan Mayan language (section 2). Kaqchikel has a number of properties, both grammatical and sociolinguistic, which differentiate it from most of the majority languages typically encountered in the speech perception literature. As discussed in section 9, our study replicates some past results regarding the influence of segment-level distributional statistics on speech perception, and in doing supports the generality of such effects across different linguistic populations.

We investigate these issues using an AX discrimination study of the stop consonants of Kaqchikel (section 3). Our emphasis in this paper is the influence of linguistic experience on speech perception in a lesser-studied language. For reasons of space we do not discuss specific patterns of pairwise consonant confusion in detail here.

2 Kaqchikel

Kaqchikel is a K'ichean-branch Mayan language spoken by over half a million people in southern Guatemala (Fig. 1; Richards 2003, Maxwell & Hill 2010, Fischer & Brown 1996: fn.3). Like all Mayan languages, Kaqchikel has a phonemic contrast between plain voiceless plosives (/p t k q t͡s t͡ʃ/) and 'glottalized' plosives at corresponding places of articulation (implosive /ɓ/, ejective /tʔ kʔ qʔ t͡sʔ t͡ʃʔ/, and /ʔ/) (Table 1; Campbell 1977, Chacach Cutzal 1990, Cojtí Macario & Lopez 1990, García Matzar et al. 1999, Majzul et al. 2000, Brown et al. 2010, Bennett 2016, etc.).

Ejectives and implosives are reasonably uncommon cross-linguistically: Maddieson's (2009) typological survey of 566 languages finds that 151 (27%) have either ejectives or implosives in their consonant inventories. Bennett et al. (submitted) find that the ejectives of Kaqchikel closely resemble the 'slack' ejectives described by Lindau (1984) and Kingston (1984, 2005): they are characteristically produced with short VOTs and weak release bursts, and cause creaky voice on adjacent voiced segments. Ejectives are sometimes realized with glottal closure following the oral release burst: this difference in release quality, along with creakiness in adjacent segments, seems to be a reliable cue to the plain∼ejective distinction in Kaqchikel. Implosive /ɓ/ usually lacks a release burst entirely, being ingressive, and also conditions creaky voice on neighboring sounds (Majzul et al. 2000, Bennett 2016). Common

Figure 1: Map of Guatemala showing the four administrative departments in which Kaqchikel is most widely spoken as a community language (from east to west, these are the departments of Guatemala, Sacatepéquez, Chimaltenango, and Sololá) (Richards 2003, Brown et al. 2010, Maxwell & Hill 2010)

realizations of /ɓ/ include [ɓ̰ ɓ̥ ɓ]; less common realizations include [pˀ w̰ ʔ]. The phonetic realization of /qˀ/ is typically either [qˀ] or [ɢ̥ˀ]. These findings are all in line with past descriptions of Kaqchikel and other Eastern Mayan languages (e.g. DuBois 1981, England 1983, Kingston 1984, Larsen 1988, Pinkerton 1986, Russell 1997, Barrett 1999, Bennett 2010). Since not much previous research has been done on the perception of glottalized consonants in any language, and none at all on Mayan languages, we do not have any prior expectations as to how discriminable plain∼glottalized contrasts might be in Kaqchikel (on the perception of glottalized consonants outside of Mayan, see Fre Woldu 1985, Wright et al. 2002, Rose & King 2007, Gallagher 2010, 2011, 2012, 2014).

The morphological system of Kaqchikel is moderately agglutinating, especially with verbs (see Chacach Cutzal 1990, Kaufman 1990, García Matzar et al. 1999, Brown et al. 2010, Coon 2016). Across lexical categories, the prefixal field is mostly reserved for inflectional affixes marking aspect and person/number agreement, while the suffixal field is composed of

	Bilabial	**Dental/ alveolar**	**Post-alveolar**	**Velar**	**Uvular**	**Glottal**
Stop	p ɓ̥	t tˀ		k kˀ	q qˀ∼ɢ̥ˀ	ʔ
Affricate		t͡s t͡sˀ	t͡ʃ t͡ʃˀ			
Fricative		s	ʃ	x ∼ χ		
Nasal	m	n				
Semivowel	w		j			
Liquid		l r				

Table 1: The phonemic consonants of Kaqchikel

derivational affixes (1,2) (the adjectival root *ch'u'j* /t͡ʃʔuʔχ/ 'crazy' is in bold).[2]

(1) x-i-b'e-ki-**ch'uj**-ir-isa-j
ASP-1SG.ABS-DIR-3PL.ERG-crazy-INCH-CAUSE-TRANS
'they went somewhere to drive me crazy'

(2) qa-**ch'uj**-ir-isa-x-ik
1PL.ERG-crazy-INCH-CAUSE-PASS-NOM
'our being driven crazy'

All modern Mayan writing systems are alphabetic in nature. Literacy in Kaqchikel is currently quite low, in part because written materials are not widely available (on the history of literature and literacy in Kaqchikel, see Maxwell & Hill 2010; on literacy in Mayan languages more generally, see England 1996, 2003, Brody 2004, and references there). While standard orthographies exist for most Mayan languages (Kaufman 2003, Bennett et al. 2016), there is a substantial amount of variability in writing conventions across speakers (England 1996, 2003, Brown et al. 2010). This variation reflects, at least in part, the extensive dialect variation found for those Mayan languages which (like Kaqchikel) are spoken over a wide geographical area (e.g. Majzul et al. 2000, Richards 2003, Brown et al. 2010, Maxwell & Hill 2010).

3 Perception study: AX task

To investigate the role that lexical and acoustic experience play in speech perception in Kaqchikel, we carried out a simple AX discrimination task investigating perceptual similarity among stop consonants.

3.1 Method

In this study, Kaqchikel speakers listened to pairs of [CV] or [VC] syllables over headphones. We will sometimes refer to the [CV] condition as the 'Onset' condition, and the [VC] condition as the 'Coda' condition. The vowels in a given syllable pair were always identical, but the consonants could be either identical or different. Upon listening to each pair of syllables, the participants were asked to respond SAME or DIFFERENT on a button box. Our underlying assumption is that incorrect SAME responses to syllables containing different consonants indicates perceptual similarity between $[C_1]\sim[C_2]$ pairs. Further details of the methodology are outlined below.

3.1.1 Participants

45 experimental participants were recruited in Patzicía, Guatemala (Fig. 1) by one of the authors (Ajsivinac), who is himself a native speaker of the Patzicía variety of Kaqchikel. These participants all have self-reported native-level fluency in Kaqchikel, a fact further

[2]Glossing conventions follow the Leipzig Glossing Rules (`https://www.eva.mpg.de/lingua/resources/glossing-rules.php`) and the Mayan-specific conventions set out in Bennett et al. (2016).

confirmed by co-author Ajsivinac during conversations before and after the experimental sessions. As is typically the case in Guatemala, most of these participants were also fully bilingual in Spanish. Kaqchikel is nonetheless the primary medium of communication in Patzicía, and the language most likely spoken by our participants at home and in many public contexts. All communication before, during, and after experimental sessions was conducted in Kaqchikel (the first author, Bennett, is a second-language speaker of Kaqchikel with conversational abilities).

Participants completed a consent form and were given 200 Guatemalan quetzals (≈$27.25) for their participation. All participants were born in the department of Chimaltenango (Fig. 1), where they also resided at the time of the study. Forty-one participants were born in the town of Patzicía itself, and 43 were living there at the time of the study. Ages ranged from 18 to 79 years old (Mean: 29, SD: 12.3). Thirteen male and 32 female speakers participated in the study (M:F ratio: 0.41). The skew toward female participants is typical of fieldwork in Guatemala, as women typically have greater flexibilty during the workday than men. One participant was excluded from analysis for failure to complete the study.

All experimental sessions were carried out in Patzicía, Guatemala (Fig. 1), in a quiet room made available to the authors for the purposes of the study. Each session took about 35 minutes to complete.

3.1.2 Stimulus design

Recording and pre-processing The stimuli used in this study were recorded by a male native speaker of Patzicía Kaqchikel (co-author Ajsivinac). The stimuli were recorded in [pVC] and [CVp] frames. These frames were chosen for several reasons. First, the dominant shape of root morphemes in Kaqchikel (as in other Mayan languages) is /CVC/: there are few content words of the shape /CV/ or /VC/ (e.g. Bennett 2016 and references there). Recording the materials as [pVC]/[CVp] helps minimize any phonetic artifacts which might result from recording materials that are not native-like in form. Furthermore, /VC/ roots are subject to consonant epenthesis in Kaqchikel, being realized as [ʔVC] in isolation, which makes it effectively impossible to record simple [VC] syllables. A plain consonant (/p/) was chosen as the frame consonant, rather than an ejective or implosive, to avoid any coarticulatory glottalization on the vowel (see Bennett 2016, Bennett et al. submitted). The frames [pVC] and [CVp] were recorded for all combinations of the vowels /a i o/ matched with each of the 22 phonemic consonants of Kaqchikel (Table 1).

The stimuli were presented for recording in random order, using an HTML platform (El Hattab 2016). For each stimulus, the speaker was asked to produce 3 repetitions with roughly even intonation. Only the best repetition for each stimulus was selected for further processing and presentation. Each recording was first manually annotated on the segmental level using the acoustic analysis program Praat (Boersma & Weenink 2016). A new set of stimuli was then extracted at these segmental boundaries, with the frame consonant /p/ excluded. The exclusion of /p/ was determined on the basis of the waveform, spectrogram, and listener audition (by co-author Bennett). Following Cutler et al. (2004), the stimuli were then amplitude-normalized with respect to the rms amplitude of the vowel (set at 60dB).

Embedding in noise In order to increase the likelihood of response errors in our study, we masked the stimuli in speech-shaped noise at a signal-to-noise ratio (SNR) of 0dB. After amplitude normalization, each stimulus was padded with 250ms of preceding silence, and 250ms of following silence. The padded stimuli were then embedded in speech-shaped noise at 0dB SNR (on our choice of SNR, see Meyer et al. 2013, Tang 2015: Ch.3.7).[3]

Stimulus pairs As noted above, participants in this study listened to [CV] or [VC] syllables presented in pairs: [C_AV]~[C_BV] or [VC_A]~[VC_B]. The perception study was designed to focus on perceptual confusion between the stops /p t k q ɓ tˀ kˀ qˀ ʔ/. Our TARGET PAIRS were pairs of [CV] or [VC] syllables in which both consonants belonged to this set of stops. All other consonants of Kaqchikel were included as fillers in this study, so that participants also heard many filler pairs in which at least one consonant was not a stop. In each [CV]/[VC] syllable the vowel was always one of /a i u/, and vowel quality was always matched between syllables presented in a pair.

There were 270 distinct target pairs in our study, ignoring the order of presentation of the items in each pair. This included 54 SAME target pairs (9 stops × 3 vowels × 2 syllable templates) and 216 DIFFERENT target pairs ($_9C_2$ (=36) consonant pairs × 3 vowels × 2 syllable templates). There were 1248 additional filler pairs, which reflect all possible combinations of non-stop consonants (n=13) with other consonants (n=22), across 3 vowel and 2 syllable contexts. The ratio of same:different trials in the study was set at 3:4 (including both filler and target pairs).

In order to keep the experiment to a reasonable length, we divided our target pairs into 30 different lists. For each list, we randomly sampled 72 DIFFERENT target pairs, and 54 SAME target pairs. Sampling of SAME trials was done with replacement, so that each list could contain multiple instances of a given SAME pair (up to a maximum of 3 repetitions).

Within each list we also included 74 filler items composed of consonant pairings that included at least one non-stop consonant. These 74 filler items were sampled with the same 3:4 ratio used to balance same:different trials for the target pairs (32 SAME fillers, 42 DIFFERENT fillers). This resulted in 200 trials per list. The 45 participants were assigned a list in order: since there were only 30 stimulus lists, the first 15 lists were assigned to two participants each, and the remaining 15 lists assigned to just one participant each. The order of presentation for the pairs in each list was randomized across participants.

3.1.3 Stimulus presentation

Presentation of the stimuli and logging of participant responses was carried out with a script written in PSYCHOPY (Version 1.82.01; Peirce 2007) and excecuted on a laptop computer. As noted above, this script assigned each participant to one of 30 stimulus lists, and automatically randomized presentation of stimuli within each list.

[3]The speech-shaped noise was generated from a four-hour acoustic corpus of spontaneous spoken Kaqchikel (section 4.1), using a PRAAT script in the library `praat-semiauto` (McCloy 2014). This Praat script took a directory of `.wav` files extracted from the spoken corpus and generated a Gaussian noise file which was spectrally shaped to match the long-term average spectrum of that corpus (essentially following Quené & van Delft 2010).

Prior to the beginning of the experiment, participants were told that they would be listening to a series of syllable pairs, and that they would have to respond as to whether they thought the syllables in each pair were the same or different. They were also told that the stimuli would be embedded in noise of some kind, making them difficult to hear, and that they should not expect the syllables to correspond to actual words of Kaqchikel in most cases (see section 5.3). This information was provided because pilot testing suggested that the presence of noise and the nonce-word status of the stimuli might be potentially confusing to some participants.

On each trial, participants were first presented with a cross in the center of the screen, lasting 500ms. The screen then changed to a display showing a green box on the left side of the screen (corresponding to SAME responses) and a red box on the right (corresponding to DIFFERENT responses). Simultaneous with this change in the display, the first member of the stimulus pair for that trial began to play over headphones (Shure SRH 440 over-ear headphones, connected to the computer via an external FiiO E10 USB preamp set at a fixed level across sessions). The order of presentation of the two syllables in a stimulus pair was randomized on each trial.

Upon hearing each stimulus pair, participants responded as to whether they thought the two syllables were identical or different, using a PST Serial Response Box attached to the laptop. SAME responses were entered with the leftmost key, and DIFFERENT responses with the rightmost key; the position of the response keys was not counter-balanced across participants.

Participants were instructed to respond as quickly and as accurately as possible. Participants could take as long as they liked to respond, but trials taking longer than 10 seconds were followed with a reminder to respond as quickly as possible (a yellow warning sign symbol). Even without significant time pressure, participants responded in under one second on most trials (mean RT = 854ms, median RT = 664ms). Ten practice trials were completed prior to the actual experiment; these practice trials always involved syllable pairs which were not included in the test list for that speaker's session.

After each participant was comfortable with the practice items, they began the actual experiment. The inter-stimulus interval (ISI) between the two stimuli in each pair was set to 300ms. The inter-trial interval was set at 1500ms (1000ms of blank screen followed by the 500ms cross fixation at the beginning of each trial). Participants were permitted to take a break after every 40 trials. The stimuli were presented at a fixed volume across trials, set at a comfortable level for each listener.

To verify that participants had completed the task as requested, we computed d′ scores (Macmillan & Creelman 2005) for perceptual confusions between each pair of stop consonants, collapsing comparisons across all participants. d′ is related to accuracy, but controls for response bias, in particular the tendency to favor one of the two responses regardless of what the stimulus is. The mean d′ score for comparisons in the Onset condition was 1.62, and the mean d′ score for comparisons in the Coda condition was 1.82.

We believe that these are reasonably good d′ values, given that our participants had limited or no prior experience as experimental participants and were not necessarily accustomed to using a computer.[4] We conclude that the participants in this study completed the task

[4] d′ scores are z-scores, so a d′ of 1 would be obtained for a participant who responded with roughly

as requested.[5]

A 9-by-9 plot summarizing the d′ scores for all target stop pairs, collapsed across vowel context and syllable position, is provided as an appendix (Appendix D).

3.1.4 ISI length and processing mode

The ISI used in this study can be estimated in at least two ways. The shortest estimate would be 300ms, the length of the silent interval between the two syllables in a given stimulus pair. If we also include the noise padding present at the beginning and end of each stimulus, then the ISI would instead be 800ms in length (250ms of noise padding before/after each syllable + 300ms silence between items). We note these values because the duration of the ISI in AX discrimination and related tasks is known to affect the way in which listeners process auditory stimuli (Pisoni 1973, 1975, Pisoni & Tash 1974, Fox 1984, Werker & Tees 1984, Werker & Logan 1985, Kingston 2005, Babel & Johnson 2010, McGuire 2010, Kingston et al. 2016). Shorter ISIs, particularly those without intervening noise, tend to favor a more acoustically-oriented mode of speech processing which does not necessarily engage the phonemic and lexical levels of speech encoding (i.e. short ISIs encourage a 'prelinguistic' mode of listening; see section 8). Longer ISIs, typically at 500ms or above, seem to condition responses which are more strongly affected by the phonemic and lexical structure of the listener's native language (i.e. a 'linguistic' mode of speech processing). We mention these considerations because an ISI of 800ms may have facilitated a linguistically-oriented mode of listening, a fact which is relevant given our goal of linking perceptual confusions to statistical facts about words and segments in Kaqchikel.

4 Two corpora for Kaqchikel: Assessing the effect of statistical and acoustic factors on speech perception

The overarching goal of this study was to assess the extent to which prior linguistic experience with Kaqchikel might affect consonant discrimination in a perceptual task. To that end, we examined acoustic, segmental, and word-level factors which could play a role in conditioning consonant confusions. Doing so necessitated the development of two corpora for Kaqchikel, which are described in the following sections.

4.1 Spoken Kaqchikel: The Sololá corpus

4.1.1 Corpus collection

The Sololá corpus is a collection of audio recordings of spontaneous spoken Kaqchikel. This corpus consists of recordings made by two of the authors in Sololá, Guatemala (Fig. 1) in

69% accuracy on both same and different trials, and a d′ of 1.5 would be obtained for a participant who responded with a bit more than 77% accuracy on both same and different trials.

[5]In addition to computing d′ scores over stop combinations, we computed a d′ score for each participant. One participant had a very low d′ score (0.047), more than 3 standard deviations from the mean d′ score across participants. We re-ran our best statistical model (3) with this participant's results excluded, and the model statistics remained virtually the same.

2013 (Bennett & Ajsivinac Sian In preparation, c). Sixteen speakers of the Sololá variety of Kaqchikel contributed to this corpus and shared short, spontaneous narratives of their own choosing for the recording.

Fifteen (out of 16) of the speakers were born in the department of Sololá. The remaining speaker was born in the department of Sacatepéquez, to the east of Sololá. As of 2013, the speakers were all living in the department of Sololá, with six living in the city of Sololá, and ten in other towns. Six of these speakers were male, and 10 female; their ages ranged from 19-84 years old (mean = 33 years, median = 28 years, SD = 15.4). The speakers all had self-reported native-level fluency in Kaqchikel, a fact further confirmed by co-author Ajsivinac during conversations before and after the recording sessions. Most speakers reported using Kaqchikel as the primary language of communication at home.

All speakers were recorded using a headset microphone (Audio-Technica ATM73a) and solid-state portable recorder (Fostex FR-2LE), at a 48kHz sampling rate with 24 bit quantization. The recordings were subsequently downsampled to 16kHz for forced alignment and acoustic analysis (see below).

4.1.2 Corpus processing

In total, the corpus amounts to about 4 hours of recorded speech (≈40,000 word tokens). The entire corpus was transcribed orthographically by one of the authors, a native speaker of Kaqchikel (Ajsivinac). We took a subset of this corpus, consisting of approximately 3.5 minutes of audio per speaker (about 50 minutes in total, consisting of 5218 word tokens and 2754 stop consonant tokens), and annotated it phonetically using forced alignment tools. First, the transcriptions in this subset of the corpus were double-checked by another author (Bennett, a trained phonologist and phonetician, as well as an L2 speaker of Kaqchikel with conversational-level abilities). The orthographic transcriptions for this portion of the corpus were then converted into a surface phonetic transcription with a suite of Python scripts (`http://www.python.org/`) implementing grapheme-to-phoneme conversion and several major allophonic rules (see DiCanio et al. 2013 for discussion).[6]

These phonetic transcriptions, and their associated audio, were then submitted to segment-level forced alignment using the ProsodyLab-Aligner (`http://prosodylab.org/tools/aligner/`; Gorman et al. 2011). Forced alignment is a computational technique for semi-automatically time-aligning audio files with a corresponding transcription. The ProsodyLab-Aligner takes as its input an audio file with an associated sentence-level transcription, and produces a time-aligned Praat TextGrid with annotations at the word and segment levels. An alignment model was first trained on the 50 minute sub-corpus (3 training rounds of 1000 epochs each), then applied to that same corpus to generate the segmental annotations. A total of 2754 stops were annotated at the segmental level using this technique. Alignments were visually-inspected by one of the authors (Bennett, a trained phonologist and phonetician), but not hand-corrected for the purposes of this analysis (see DiCanio et al. 2013 on the distribution of error types in forced alignment).[7]

[6]This suite of Python scripts is currently not available, as they are being developed as part of another ongoing project.

[7]As a rough assessment of the accuracy of our forced alignment model, we hand-corrected a subset of the TextGrids produced by forced alignment, and compared them to the original, automatically aligned output.

4.1.3 Corpus criticism

There are several advantages to using a spoken corpus of this type for acoustic analysis and psycholinguistic research. First, the Sololá corpus is a corpus of spontaneous speech, and is therefore more naturalistic, and more representative of everyday Kaqchikel speech, than a corpus of read or elicited materials. Second, the materials in such a corpus—which include stories and folktales that are traditionally told in the Sololá region—may be of greater interest to the Kaqchikel language community than recordings of isolated wordlists or prompted sentences.

There are also potential drawbacks to using a corpus of this type. While the Sololá corpus has the advantage of being naturalistic and thus more ecologically valid than certain other types of audio corpora, the content of the recordings is not controlled in any way (see Xu 2010 for discussion). As a consequence, data sparsity issues emerge with respect to certain phonetic and phonological structures. For example, ejective /tʔ/ is quite rare in our data (<1% of stops). (This is to be expected, as ejective /tʔ/ is known to have low type and token frequencies in Mayan languages; England 2001, Bennett 2016.) In any case, the paucity of /tʔ/ tokens in the corpus clearly precludes any strong conclusions about the properties of this sound. Additionally, there are relatively few glottalized stops in non-prevocalic (≈ coda) position in our corpus (<5% of all stops). This owes in part to the fact that most stops in our corpus, regardless of laryngeal state, occur in pre-vocalic position (85%).

Although the Sololá corpus is a corpus of spontaneous speech, it is also a corpus of monologues rather than dialogues. As such, the speech genre of the corpus may be less than fully naturalistic, and may further show the effects of stereotyped or ritualistic speech patterns associated with storytelling in Kaqchikel. Nonetheless, we believe that the size and composition of this corpus is appropriate for drawing at least some initial conclusions about the phonetic structure of everyday Kaqchikel speech.

4.2 Written corpus

4.2.1 Corpus collection

One goal of this paper is to explore how the statistical structure of Kaqchikel—both in the lexicon (i.e. the vocabulary), and in actual spoken or written usage—might influence speech perception. To answer this question we needed a reasonably large corpus of written Kaqchikel over which segmental and word-level statistics could be calculated. To the best of our knowledge there are no structured corpora of written Kaqchikel currently available (apart from dictionaries like Macario et al. 1998, Majzul 2007), and certainly none that are in a digitized, searchable form. It was therefore necessary to construct a novel, digitized written corpus of Kaqchikel in order to assess the statistical patterning of words and segments in the language.

On average, stop consonants were well-identified by our alignment model: out of 428 stops, the median alignment error was 10ms (mean = 16ms) (see also DiCanio et al. 2013). Further, 25% of alignments agreed to the exact millisecond, and 86% of alignment errors were 20ms or shorter in size. These errors appear to be more-or-less evenly distributed across stop types: consequently, alignment errors are unlikely to have skewed our measures of perceptual similarity in any particular direction. We thank Andrea Maynard for carefully hand-correcting these TextGrids.

Our corpus is constructed from religious texts, spoken transcripts, government documents, medical handbooks, and other educational books written in Kaqchikel—essentially all the materials we could find that were already digitized or in an easily digitizable format. The corpus contains approximately 0.7 million word tokens (around 30,000 word types).

The corpus underwent further processing and cleaning before being used to calculate word- and segment-level corpus measures for Kaqchikel. Details on our processing and cleaning methods are given in Appendix A.

4.2.2 Corpus criticism

Corpus size and composition Modern corpora of majority languages like English are quite large, on the order of hundreds of millions of word tokens in size (e.g. the Subtlex-UK corpus, 201 million words, van Heuven et al. 2014). Spoken corpora tend to be smaller, but still typically contain several million words (e.g. the Corpus of Spontaneous Japanese, 7 million words, Maekawa 2003). Developing corpora of this size is simply not feasible for under-resourced languages like Kaqchikel, which may lack large quantities of written text (particularly digitized text), as well as the economic infrastructure needed to support the collection and annotation of large corpora.

For this reason, in compiling our written corpus we drew on any and all written Kaqchikel texts that we could find. We purposefully excluded dictionaries and collections of neologisms from the corpus because these sources are likely to contain words which are not familiar to most Kaqchikel speakers.

In several respects, our written corpus is far from ideal. First, the corpus is relatively small, containing only ≈0.7 million word tokens. It has been argued that a corpus of 16 million word tokens or more is needed for calculating stable estimates of the statistical properties of low frequency words (Brysbaert & New 2009).

Second, our corpus contains a mix of both spoken transcripts and written sources. Ideally we would make use of a corpus consisting exclusively of spoken transcripts, given that most Kaqchikel speakers are not literate in the language, or otherwise have limited experience reading in Kaqchikel. Even for majority languages with higher literacy rates, it has been argued that spoken corpora are more representative of speakers' actual linguistic experience than written corpora (Brysbaert & New 2009, Keuleers et al. 2012).

Additionally, it is important to recognize that the Kaqchikel orthography is only semi-standardized, and orthographic practices vary across dialects and speakers of the language (Brown et al. 2010: 3-4). For example, our corpus contains both *nb'än* and *nub'än* as forms of '(s)he does it', reflecting the fact that some dialects omit the 3SG.ERG marker *-u-* in particular morphological contexts (Majzul et al. 2000: 69-70).

Genre The *representativeness* of a corpus refers to how closely a corpus reflects actual language use in a particular population (e.g. Atkins et al. 1992, Biber 1993). One measure of representativeness is the extent to which the texts and genres included in a corpus correspond to the kinds of texts (or linguistic interactions) that speakers in the target population typically engage with.

The written Kaqchikel corpus described here is not balanced by genre, nor is it particularly speech-like with respect to the thematic content of the materials that it includes. To

get a rough sense of how far the corpus deviates from naturalistic speech, we used the transitional probabilities between words in the corpus to create a trigram Markov-chain language model. Using this Markov-chain language model, we stochastically generated ('babbled') some random samples of Kaqchikel. One such sample is shown below.

A sample of Markov-chain Kaqchikel

> ri taq Mechanpomal moloj: achoq pa ruwi' yesamäj. K'ïy mul nqak'axaj nkib'ij chi ri xaqixaq nuqasaj ri k'atän jub'a'. K'o b'ey chuqa' nq'axon nchulun o taq nsinan. K'ïy b'ey man ntane' ta ri retal nuya' chi ke ronojel ri qamolojri'ïl. Richin nawetamaj más República Democrática del Congo Ruanda Jun peraj chi re ri raqän ya' Jordán. Ri Jehová rik'in ri más ütz chuqa' man ütz ta yojch'on rik'in jun winäq...

Loose English translation of the Markov-chain Kaqchikel sample

> the Mechanpomal group: on top of what do they work. Many times we listen to what they say about wormwood which lowers the heat a little. There are times too it hurts when he urinates or has sexual relations. Many times it doesn't stop, the sign it gives to all of our organizations. In order for you to know more Democratic Republic of the Congo Rwanda A shawl for the river of Jordan. Jehova is with the best and it isn't good that we talk to a man...

It is clear from the sample that the written corpus is not particularly speech-like, although it does contain a good range of lexical items covering the topics of religion, geography and agriculture. Despite the fact that the genres represented by our corpus diverge somewhat from everyday speech, Tang et al. (In preparation) show that word frequencies estimated from this written corpus can be used to predict the duration of words in our corpus of spontaneous spoken Kaqchikel (section 4.1; see C. E. Wright 1979 for the classic finding that word frequency and word duration are correlated in English). We take this result as indirect evidence that our written corpus roughly approximates the lexical structure of Kaqchikel as it is actually spoken.

For present purposes, the question is whether measures like functional load or phoneme frequency can be reliably estimated from this corpus. Work in progress (Tang et al. 2015) suggests that estimates of these measures are stable even over small sub-samples of this corpus (e.g. 20,000 words; see too Gasser & Bowern 2014, Macklin-Cordes & Round 2015, Dockum & Campbell-Taylor 2017). As such, we believe that our corpus is indeed of sufficient size for the estimation of these measures.

5 Predictions and model design

In this section we consider how acoustic factors, word-level statistics, and segmental statistics might interact with speech perception in Kaqchikel. We also describe the basic modeling procedure we used to test whether these predictions were borne out in our results (section 6).

5.1 Statistical model

We analyzed participant accuracy on each trial of the AX discrimination task with a mixed effects logistic regression in R (R Development Core Team 2013), using the `glmer` function in the `lme4` library (Bates et al. 2011). Recall that each trial could either contain two identical stimuli (the SAME condition), or two different stimuli (the DIFFERENT condition). We interpreted incorrect responses on DIFFERENT trials as evidence that a given pair of stimuli was perceptually similar (having been mistaken as identical). SAME trials are not similarly informative regarding perceptual confusion; we therefore analyzed only the accuracy of participant response on DIFFERENT trials.[8]

5.2 Acoustic similarity

One of our main expectations is that greater acoustic similarity between a pair of syllables should predict greater perceptual similarity between those syllables (e.g. Dubno & Levitt 1981). Two acoustic similarity measures were considered. The first measure is STIMULUS SIMILARITY—the raw acoustic similarity of the stimuli themselves. We expected stimulus similarity to have a substantial effect on stimulus discrimination in our study.

The second measure is CATEGORY SIMILARITY—the similarity of two phoneme categories based on *prior phonetic experience*. Following a large body of work in Exemplar Theory, we assume that phonemic categories are associated with episodic memory traces (or exemplars), which are phonetically-rich representations of that category as previously encountered on specific occasions in actual speech (Goldinger 1996, 1998, Pierrehumbert 2001, 2002, Wedel 2004, Gahl & Yu 2006 and references there). On this view, the category similarity of two phonemes can be conceived of as the extent to which their exemplar clouds show overlap in phonetic space (see also Yu 2011).

Figures 2 and 3 illustrate the importance of distinguishing stimulus similarity and category similarity. These figures show two hypothetical exemplar clouds for the phonemes /k/ and /p/ over some acoustic dimension(s) (say, VOT and burst intensity). The two clouds represent a collection of prior phonetic experiences that the listener has associated with each phoneme. In the context of our study, the two dots represent two stimuli presented to the listener for discrimination (say [ka] and [pa]).

In Figure 2, the exemplar clouds are non-overlapping (i.e. VOT values for /k/ and /p/ are typically quite distinct). As a consequence, listeners would likely conclude that the two stimuli (the dots) belong to different categories. In Figure 3, the clouds are substantially more overlapped, with the two stimuli falling in the overlapping region. This overlap between the two categories increases the level of uncertainty for the listener, making it more difficult to determine whether the two stimuli belong to different phonemic categories. To the extent that overlap along particular phonetic dimensions makes listeners less likely to rely on those dimensions for category discrimination (e.g. Holt & Lotto 2006), category overlap may

[8] SAME trials are often taken into account in statistical analyses of discriminability based on d′ (Macmillan & Creelman 2005), as a way of controlling for individual response biases. In our model, response biases are captured by including PARTICIPANT as a random effect in the linear regression.

A 9-by-9 plot summarizing the d′ scores for all target stop pairs, collapsed across vowel context and syllable position, is provided as an appendix (Appendix D).

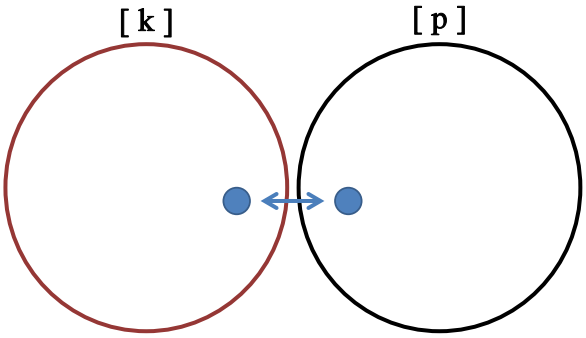

Figure 2: Non-overlapping exemplar clouds

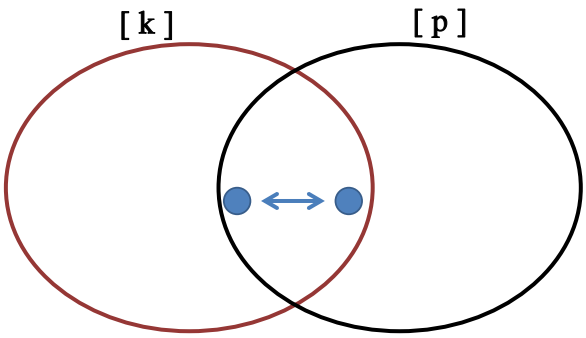

Figure 3: Overlapping exemplar clouds

influence discrimination even for stimuli which are acoustically unambiguous (i.e. in non-overlapping regions of Fig. 3), by reducing overall sensitivity to certain potential cues to consonant identity.

5.2.1 Stimulus similarity

It is intuitively clear that higher levels of acoustic similarity between stimuli should lead to higher rates of confusion between those stimuli in a discrimination task (e.g. Dubno & Levitt 1981, Redford & Diehl 1999, Hall & Hume submitted and many others). To evaluate the importance of other factors in this study—particularly those related to prior phonetic experience (category similarity)—stimulus similarity must therefore be included as a control predictor.

To capture the acoustic similarity between the stimulus pairs in each trial, an acoustic distance metric was applied to each stimulus pair (after embedding the stimuli in noise). Such a metric should allow us to capture the raw acoustic information that could be used by the listeners to perform the AX task, even without accessing higher-level perceptual, phonemic, or lexical processing. Our acoustic distance metric was calculated using PHONOLOGICAL CORPUSTOOLS (Hall et al. 2015). First, the waveform of each stimulus was transformed into mel-frequency cepstrum coefficents (MFCCs) (Mielke 2012), a common re-representation of the acoustic signal used widely in speech recognition research. The number of MFCCs was set to 12, as this allows the model to capture speaker-independent information about acoustic similarity (see `http://corpustools.readthedocs.io/en/latest/acoustic_similarity.html`). Dynamic Time Warping (DTW), another common speech processing technique, was used to compute an explicit distance metric on the basis of the MFCC-transformed stimuli (Sakoe & Chiba 1971, Mielke 2012).

Our statistical model included a fixed effect for STIMULUS SIMILARITY, representing the acoustic distance between two stimuli according to this DTW metric. This predictor was z-score normalized using the `scale()` function in R.

5.2.2 Category similarity

To calculate the acoustic similarity of two stops at the phonemic level (category similarity), we computed DTW over pairs of stops as they occur in our acoustic corpus (section 4.1). We

further limited our comparisons to stop consonants occuring in similar environments, since the confusability of any given pair of stops may depend on the phonotactic and prosodic context (e.g. Chang et al. 2001, Cutler et al. 2004).

1. Using our acoustic corpus, we identified all instances of /p t k q ɓ tˀ kˀ qˀ/. These were divided into two groups: (i) pre-vocalic ([C̲V]); and (ii) post-vocalic, but non-prevocalic ([VC̲(C/#)]).
2. Using the segmentation provided by forced alignment (section 4.1), the waveforms corresponding to each target stop consonant and the vowel adjacent to it were extracted individually. This gave two sets of waveforms corresponding to /p t k q ɓ tˀ kˀ qˀ/ in [CV] and [VC] transitions.
3. These waveforms were further divided into subsets on the basis of the vowel, matching [CV] and [VC] waveforms according to the quality and stress profile of the vowel. For example, [ˈke] could be compared with [ˈte], but not with [te], [ˈti], [ˈkɛ], or [ˈek].
4. Within each matched [CV] or [VC] subset, we computed an acoustic distance measure (DTW) between all pairs of waveforms within that set which contained different stop consonants. For example, if [ˈke] occurred twice in the corpus, and [ˈte] occurred three times, we would compute six pairwise acoustic distance measures: $[\text{ˈke}]_1 \sim [\text{ˈte}]_1$, $[\text{ˈke}]_1 \sim [\text{ˈte}]_2$, $[\text{ˈke}]_1 \sim [\text{ˈte}]_3$; and $[\text{ˈke}]_2 \sim [\text{ˈte}]_1$, $[\text{ˈke}]_2 \sim [\text{ˈte}]_2$, $[\text{ˈke}]_2 \sim [\text{ˈte}]_3$.
5. The outcome of this procedure is a set of acoustic distances between tokens of the stop categories /p t k q ɓ tˀ kˀ qˀ/, grouped according to their syllabic context (onset/coda) and the properties of the preceding/following vowel. As an aggregate measure of category similarity, we took the mean and standard deviation of these values for each pair of stops.

These measures of category similarity were then used as predictors in our analysis of perceptual similarity (section 6).[9]

Under the assumption that each stop token in the corpus counts as an exemplar, and exemplars are clustered together in clouds according to their category membership, the mean category distance between two stops can be interpreted as the distance between the two centroids of the exemplar clouds associated with each phoneme. The standard deviation of the distances between tokens is a measure of how *consistently* different the two categories are, across contexts and repetitions. These measures are logically and practically independent of each other. For instance, /ɓ/∼/p/ and /$\widehat{\text{tʃ}}$ˀ/∼/qˀ/ have similar mean acoustic distances (51.04 and 51.74 respectively), but the standard deviations of the acoustic distances associated with each category are rather different (8.80 and 5.39 respectively). Since the separation between category means and the variance around those means might both matter for the overall separation of two phonemic categories in the acoustic space, we treated both measures of category similarity as predictors in our analysis of the perception study described

[9]Our measure of category similarity (means and standard deviations of pooled DTW measurements) does not distinguish between [CV] and [VC] contexts. We initially considered computing category similarity separately for [CV] and [VC] contexts, to more closely match the experimental design of our perception study (section 3). We abandoned this approach because of the sparseness of the acoustic corpus (e.g. the rarest phoneme /tˀ/ only precedes or follows /ˈe/ and /ˈo/, and no other vowels). Coping with this problem would have required us to pool over other contextual properties, such as the quality of the adjacent vowel, and in doing so we would have ignored other perceptually relevant factors in the analysis.

above. Ultimately, only the category means proved to be a reliable predictor of perceptual similarity in our study (section 6).

Our statistical model includes two fixed effects for CATEGORY SIMILARITY between the two phonemic categories being compared on a given DIFFERENT trial: MEAN CATEGORY SIMILARITY and STANDARD DEVIATION OF CATEGORY SIMILARITY. Both predictors were z-score normalized using the `scale()` function in R.

5.3 Word- and segment-level statistical factors

The analysis of statistical effects on speech perception in Kaqchikel took into account a number of distinct segment- and word-level predictors. Only a few of these predictors made a significant contribution to predicting patterns of perceptual confusion between stops in our study (section 6). In the following section we define only those predictors which made a reliable contribution to predicting stop discrimination in our study, and leave the definition of the other, non-significant factors which we considered to Appendix C.

5.3.1 Segment-level factors

Three segment-level predictors were considered: segmental frequency, functional load, and distributional overlap. Of these, only functional load and distributional overlap emerged as significant predictors of perceptual confusions in our study.

Functional load Intuitively, FUNCTIONAL LOAD characterizes the importance of a given phonemic contrast for distinguishing words in a language. One way of defining the functional load (henceforth FL) of two phonemes in a language is to count the number of minimal pairs that are distinguished solely by the contrast between those phonemes (Martinet 1952, Kučera 1963, Hockett 1967, Surendran & Niyogi 2003, 2006). It has been argued that FL and related measures condition the probability of diachronic phoneme mergers (Wedel et al. 2013), as well as the production of phonemic contrasts (Baese-Berk & Goldrick 2009, Goldrick et al. 2013, Nelson & Wedel To appear).

Perhaps most relevant to this study, FL may also interact with the *perception* of phonemic contrasts. Graff (2012), drawing on data from 60 languages and 25 language families, argues that languages tend to use perceptually robust phoneme contrasts to distinguish minimal pairs. Consequently, there should be a positive correlation between functional load and the perceptual distinctiveness of a given phonemic contrast. Hall & Hume (submitted) and Stevenson & Zamuner (2017) show that in French, vowel pairs which have a higher functional load are also more perceptually distinct, even when other factors (such as raw acoustic similarity) are taken into account (see also Renwick 2014 for similar claims about Romanian). Relatedly, Davidson et al. (2007) found that listeners were more attentive to subtle phonetic details in an AX discrimination task (such as the presence vs. absence of schwa in clusters, [CəC]∼[CC]) when the items constituted minimal pairs.

For this study, we focused on a metric of functional load which captures the change in entropy of the lexicon following merger of a phoneme contrast. This metric is sometimes called *lexical* Δ*-entropy* (for comparison with other metrics, see footnote 13). To compute

this measure, we employed an information-theoretic method (Shannon 1948). We first calculated the ENTROPY of the Kaqchikel lexicon—a measure of uncertainty—using Equation 1 (Surendran & Niyogi 2003, 2006). For our purposes, the entropy $H(L)$ of a language measures how diverse the vocabulary is (basically the size of the lexicon), weighted by token frequency. Entropy depends on p_w, the probability of a given word w in our written corpus. Functional load (Equation 2) is measured by estimating how much the lexicon 'shrinks' when two phonemes are merged into one (i.e. the number of distinct words made homophonous, weighted by frequency). The entropy of a lexicon in which phonemes x, y have been merged, $H\left(L_{xy}\right)$, is compared to the original, non-merged lexicon $H\left(L\right)$ to yield the functional load of the x, y contrast (Equation 2). Phoneme pairs with a higher functional load should lead to a larger proportional decrease in entropy when they are merged.

$$H\left(L\right) = -\Sigma_w \; p_w \times log_2(p_w) \tag{1}$$

$$FL\left(x, y\right) = \frac{H\left(L\right) - H\left(L_{xy}\right)}{H\left(L\right)} \tag{2}$$

For the purpose of computing functional load, 'words' are defined as whole word forms, including affixes (i.e. as strings of segments separated by white space in a text; see Appendix A).

Wedel et al. (2013) found that patterns of diachronic phoneme merger were better predicted by a measure of functional load which only compares words belonging to the same lexical category (e.g. two nouns distinguished by an /A B/ phoneme contrast would contribute to the functional load of /A B/, but not a noun-verb pair). As Kaqchikel is moderately agglutinating (section 2), many words bear affixes which unambiguously indicate their part of speech (e.g. both affixes in *r-utz-il* 'its goodness' 3SG.POSS-good-NOM signal that this word is a noun). Our whole-word measure of functional load is thus probably biased toward comparing words within the same lexical category, as in Wedel et al. (2013). Unlike Wedel et al. (2013), we did not consider measures of functional load computed over lemmas (basically, uninflected stems) because a lemmatized corpus of Kaqchikel is not currently available.

A fixed effect of FUNCTIONAL LOAD was included in our statistical model, reflecting the measure of Δ-entropy described above.

Distributional overlap Recent work by Hall et al. (2014) and Hall & Hume (submitted) argues that the predictability of two phonemes across contexts contributes to their perceptual confusability. The theoretical context of this claim is one in which contrastiveness is assumed to be gradient rather than categorical: two phonemes which occur in many of the same environments are taken to be *more contrastive* (i.e. less predictable) than phonemes which mostly occur in distinct environments (Hall 2012, 2013). By hypothesis, phoneme pairs which are more contrastive (less predictable from context) are expected to be more readily discriminated.[10]

We took Jeffreys' distance (also called Jeffrey divergence; henceforth JD) as our measure of the distributional overlap (=contextual predictability) of two phonemes. JD determines

[10]This measure of gradient contrastiveness is, confusingly, sometimes also known as 'functional load' (e.g. King 1967).

the contextual probability of two phonemes based on the local segmental contexts X__Y that they occur in. JD can thus be interpreted as a measure of phonotactic similarity.[11]

JD was implemented using the *TiMBL* manual (Daelemans et al. 2009: 26). In our study, each segment type is a class, and our features are the presence and absence of segments; more specifically, we used a trigram sliding window to generate features, with the target segment being in the first, second, or third position of the trigram window. Trigram windows are commonly used to capture phonotactics in computational linguistics (e.g. Jurafsky et al. 2001) as well as phonology more broadly (e.g. Hayes & Wilson 2008). In the specific case of Kaqchikel, trigram windows are necessary to capture certain co-occurrence restrictions which hold between the two consonants in a /CVC/ root (see Bennett 2016, Bennett et al. submitted).

A major difference between our metric and other metrics (such as the entropy-based measures in Peperkamp et al. 2006 and Hall 2012) is that our contexts are defined over segments as opposed to phonological features. This decision owes in part to our own uncertainty about which phonological features are most appropriate for classifying segments in Kaqchikel, particularly in the case of laryngeal contrasts (see Bennett et al. submitted for discussion). The details of the metric are shown below in Equation 3.

$$JD(S_1, S_2) = \sum_{i}^{n} P(C_i|S_1) \times log(\frac{P(C_i|S_1)}{m}) + P(C_i|S_2) \times log(\frac{P(C_i|S_2)}{m}) \quad (3)$$

- $m = \frac{P(C_i|S_1)+P(C_i|S_2)}{2}$
- S_1 and S_2 are two phones.
- C_i is a phonotactic environment, defined with a sliding trigram window: __XY, X__Y, XY__

A fixed effect of DISTRIBUTIONAL OVERLAP was included in our statistical model, reflecting this measure of JD.[12]

5.3.2 Word-level factors

In addition to the variables mentioned thus far, which were all part of the experimental design, a number of nuisance variables were included in the analysis of consonant confusions: wordhood, word frequency, neighborhood density, average neighborhood frequency, and bigram frequency. Our study was not designed to test the effect of these factors, but we

[11] JD is a symmetric variant of Kullback-Leibner distance (Kullback & Leibler 1951), which has also been used to measure distributional differences across phones, for instance in research on statistical learning of allophonic alternations (Peperkamp et al. 2006, Calamaro & Jarosz 2015).

[12] We have not yet performed a stability simulation (section 4.2.2) assessing how reliably distributional overlap is estimated from corpora of different sizes. We nonetheless expect that distributional overlap can be reliably estimated from a fairly small corpus, like our corpus of Kaqchikel. Tang et al. (2015) found that estimates of functional load computed over syllable types reached stability even faster than estimates computed over word types: this likely reflects the fact that syllable types, being smaller units, are better represented in the corpus than word types. Since distributional overlap is computed over trigrams, which are similar in size to syllables, we also expect distributional overlap to be reliably estimated from a relatively small corpus.

included them in the analysis as control predictors, just in case they had an effect on our results. Of these factors, only word frequency made an appreciable contribution to predicting consonant discrimination in our study (and even then, only marginally so). We describe wordhood and word frequency here; the remaining predictors are defined in Appendix C.

Wordhood Ganong (1980) established the now classic result that listeners are more likely to identify a phonetically ambiguous segment as belonging to some phoneme P$_x$ if the categorization of that segment as P$_x$ results in an actual word of the listener's native language, and categorizing the segment as a competing phoneme P$_y$ does not (see Fox 1984, Pitt & Samuel 1993, Kingston 2005, Kingston et al. 2016 and references there).

While our stimuli did contain real words, they were only included in order to achieve balanced coverage over the consonant and vowel combinations which were the focus of this study. However, since wordhood is known to play a role in speech perception it was included as a possible predictor of perceptual confusions in our AX discrimination task.

To assess the wordhood of our stimuli, we consulted two sources: a native speaker of Kaqchikel (co-author Ajsivinac) and the headwords in two dictionaries (Macario et al. 1998, Majzul 2007). We considered a [CV] or [VC] stimulus to be a 'word' of Kaqchikel if it was identical to either a function word (e.g. the particle *k'a* /kʔa/ 'until, then, well') or a content word (e.g. *aq'* /aqʔ/ → [ʔaqʔ] 'tongue'; word-initial epenthetic glottal stops were ignored for the purposes of computing wordhood). Affixes and other bound morphemes were not considered to be words in this sense (e.g. *at-* /at-/ 2SG.ABS or *-i'* /-iʔ/ REFLEXIVE). We tailored these judgments to the Patzicía dialect: for example, *uq* [ʔuq] counted as a word because *üq* 'skirt' is pronounced as [ʔuq] (rather than historical [ʔʊq]) in the Patzicía dialect. Only 15 of our experimental items (including fillers) were actual words of Kaqchikel; the remainder (108) were non-words according to these criteria.

A fixed effect for WORDHOOD was included in our statistical model as the absolute difference of the wordhood values of two given stimuli: if both stimuli were words or both were non-words, the value was coded as 0, otherwise as 1. This predictor was z-score normalized using the `scale()` function in R.

Word frequency The effect of wordhood on phonemic categorization also obtains when the categorization of an ambiguous segment as *either* phoneme P$_x$ or phoneme P$_y$ would result in a real word, but the two resulting words differ in token frequency (de Marneffe et al. 2011). This suggests that categorization judgments can be influenced not only by the categorical word∼non-word distinction, but also by gradient differences in word frequency.

More generally, word frequency has shown to contribute to both visual and auditory word recognition, with high frequency words being recognized more accurately and more quickly than low frequency words (Howes 1957, Brown & Rubenstein 1961, Broadbent 1967, Felty et al. 2013, Tang 2015: Ch.4). Furthermore, when words are incorrectly identified in perception, the perceived word tends to have roughly the same lexical frequency as the intended word (Vitevitch 2002, Tang & Nevins 2014, Tang 2015: Ch.4, Tang & Nevins In prep). It follows that in our study, even if the stimuli (one or both) were incorrectly perceived on a given trial, the difference in word frequency between the two items could still bias the participants' responses.

Word frequency was obtained using our written corpus. We only obtained word frequency information if a stimulus was determined to be a word according the criteria described above. We used the difference in frequency between the two stimuli on a given trial as a predictor of consonant confusions; non-words were coded as having zero frequency.

A fixed effect of WORD FREQUENCY was included in our statistical model as the absolute difference of the log-transformed (base-10) word token frequencies of the stimuli in each trial, with Laplace ('add one') smoothing for frequencies of zero (prior to log-transformation; Brysbaert & Diependaele 2013). This predictor was z-score normalized using the `scale()` function in R.

6 Analysis and results

6.1 Statistical modeling

The statistical analysis began with the construction of an initial (or 'superset') model which included a large number of predictors. This initial model was then simplified through a model criticism procedure described in Appendix B. The factors included in the initial model are described below.

6.1.1 Fixed effects

As mentioned above, our initial model included fixed effects for three acoustic predictors (STIMULUS SIMILARITY, MEAN CATEGORY SIMILARITY and STANDARD DEVIATION OF CATEGORY SIMILARITY), as well as fixed effects for FUNCTIONAL LOAD, DISTRIBUTIONAL OVERLAP, WORDHOOD, and WORD FREQUENCY. Of these factors, only STIMULUS SIMILARITY, MEAN CATEGORY SIMILARITY, FUNCTIONAL LOAD, DISTRIBUTIONAL OVERLAP, and WORD FREQUENCY emerged as significant predictors of consonant discrimination in the final statistical model.

Along with these predictive factors, our initial model included fixed effects for SEGMENTAL FREQUENCY, NEIGHBORHOOD DENSITY, AVERAGE NEIGHBORHOOD FREQUENCY, and BIGRAM FREQUENCY (see Appendices B and C). These predictors were coded by log-transforming the values of the relevant measure, and taking the absolute difference of those log-transformed values for the two stimuli on each trial (Laplace smoothing was also used for WORD FREQUENCY). All five predictors were z-score transformed; none of them emerged as predictive in our final statistical model.

Response time There is a well-known trade-off between speed and accuracy in many behavioral tasks (e.g. Heitz 2014). To account for the possibility of such a tradeoff in our study, RESPONSE TIME was treated as a fixed effect predictor in the analysis of accuracy (Davidson & Martin 2013).

The response time on each trial was measured from the offset of the second stimulus (including the 250ms of noise padding following the end of the syllable itself) to the time at which the response was logged. Given that each participant might have a different baseline response speed, these response times were transformed into by-participant z-scores.

6.1.2 Random effects

Item-level random effects UNORDERED STIMULUS PAIR was treated as a random intercept, and was defined as the unordered pairing of any two DIFFERENT stimuli. As each participant heard one of 30 different lists of stimulus pairs (section 3.1.2), LIST was also included as a random intercept. The order of the two stimuli in a given trial (STIMULUS ORDER) was included as another random intercept, since the order of stimulus presentation has been reported to affect same-different discrimination judgments in some tasks (Cowan & Morse 1986, Repp & Crowder 1990, Best et al. 2001, Bundgaard-Nielsen et al. 2015, Dar et al. 2018).

Finally, the position of the target stop in each stimulus (ONSET VS. CODA) was treated as a random intercept. Phonotactic context is an important factor that influences consonant discrimination (see R. Wright 2004 for an overview). A large body of research has found that place, manner, and laryngeal features are better discriminated for prevocalic [CV] consonants (particularly stops) than for non-prevocalic [VC] consonants (e.g. Wang & Bilger 1973, Fujimura et al. 1978, Bladon 1986, Redford & Diehl 1999, Steriade 2001, 2009, Benkí 2003, Jun 2004, Tang 2015, and others; cf. Cutler et al. 2004, Meyer et al. 2013 for skeptical views). Our initial analysis of d′ found no distinction in perceptibility between onset [CV] and coda [VC] contrasts (section 3.1.3), but it still seemed prudent to include consonant position as a potential predictor of consonant confusions in this study.

Participant-level random effects PARTICIPANT was treated as a random intercept to control for inter-speaker differences in overall accuracy. In addition, by-participant random slopes for all of the word- and segment-level factors mentioned above were also included in the initial model. These by-participant random slopes were motivated by the fact that vocabulary size—which may vary across individuals—has been shown to associate with the effect of lexical factors like neighborhood size and average neighborhood frequency (Yap et al. 2015).

6.1.3 Procedure

We began with an initial, full model incorporating all of the fixed and random effects described above. This model was then simplified by a standard step-down model-selection procedure making use of the `anova()` function and likelihood ratio test provided by R. This procedure, described in greater detail in Appendix B, resulted in the final, best model in (3), where (1|F) indicates a simple random effect of factor F.

(3) Best model

ACCURACY ~ STIMULUS SIMILARITY + CATEGORY SIMILARITY (MEAN) + FUNCTIONAL LOAD + DISTRIBUTIONAL OVERLAP + WORD FREQUENCY + (1 | UNORDERED STIMULUS PAIR) + (1 | PARTICIPANT)

6.2 Statistical results

6.2.1 Unimportant factors

A number of fixed and random effects were dropped during the model selection procedure. The fixed effects which fell out of the model were CATEGORY SIMILARITY (SD), SEGMENTAL FREQUENCY, all but one of the word-level factors (WORDHOOD, NEIGHBORHOOD DENSITY, AVERAGE NEIGHBORHOOD FREQUENCY and BIGRAM FREQUENCY) and RESPONSE TIME. We suspect that CATEGORY SIMILARITY (SD) emerged as insignificant because CATEGORY SIMILARITY (MEAN) provides a better estimate of the distance between two phonemic categories: CATEGORY SIMILARITY (MEAN) reflects the distance between category centroids—a property which clearly impacts the overall similarity between two categories—while CATEGORY SIMILARITY (SD) reflects the variability in pairwise token comparisons across those categories—a property which could lead to either more or less overlap between categories depending on the shape of the variation. Like Bundgaard-Nielsen & Baker (2014) and Bundgaard-Nielsen et al. (2015), we did not find an effect of SEGMENTAL FREQUENCY. Most word-level predictors were dropped from our final model, which we interpret as evidence that word-level factors had a limited effect on discrimination accuracy in our study. This perhaps reflects the fact that listeners could carry out the task (AX discrimination) without accessing lexical items of Kaqchikel (see sections 5.3.2, 8 for more discussion). The insignificance of RESPONSE TIME further suggests that there was no meaningful speed-accuracy trade-off in this study.

The dropped random intercepts were STIMULUS ORDER, ONSET VS. CODA, and LIST. All of the by-participant random slopes for word-level factors also fell out of the final model. Unlike e.g. Bundgaard-Nielsen et al. (2015), we found no evidence that the order of presentation of the two stimuli within a pair affected participant responses. The insignificance of LIST suggests the stimuli were randomized successfully across participants, such that the distribution of stimuli within and across lists did not serve as an accidental confound. The failure to retain by-participant random slopes for word-level factors in the final model may owe to several factors: either vocabulary size was fairly homogenous across participants, or differences in vocabulary size do not have a material effect on the strength of the statistically significant predictors (segment-level FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP, and WORD FREQUENCY).

6.2.2 Explanatory factors

The significant fixed factors in the best model are reported in Table 2.

Two out of three of the acoustic similarity measures are highly significant, particularly STIMULUS SIMILARITY. Both acoustic measures have negative coefficients, meaning that the greater the acoustic similarity between two stimuli and their associated phonemic categories, the harder it is to discriminate those stimuli. Second, FUNCTIONAL LOAD has a positive coefficient, meaning that the higher the pairwise functional load of two stops, the easier it is to discriminate syllables differentiated by those stops. Third, DISTRIBUTIONAL OVERLAP has a negative coefficient, meaning that the more phonotactic environments shared by two stops, the harder it is to discriminate them (this was an unexpected finding, which we discuss in detail below). Fourth, the only remaining word-level factor, WORD FREQUENCY,

	β	SE(β)	$\lvert z\rvert$	p-value
(Intercept)	0.8042	0.1621	4.963	$< .001^{***}$
STIMULUS SIMILARITY	-1.0720	0.1151	9.316	$< .001^{*}$
CATEGORY SIMILARITY (MEAN)	-0.3876	0.1238	3.131	$< .005^{**}$
FUNCTIONAL LOAD	0.4653	0.1649	2.822	$< .005^{**}$
DISTRIBUTIONAL OVERLAP	-0.6320	0.1607	3.933	$< .001^{***}$
WORD FREQUENCY (ABS. DIFF.)	0.1848	0.1068	1.731	$.084^{n.s.}$

Table 2: Regression statistics of the fixed effects in the best model predicting response accuracy (correct: 1, incorrect: 0) in log odds space. '*' stands for $p < .05$; '**' for $p < .01$; '***' for $p < .001$; '.' for $p < .10$; $^{n.s.}$ for 'not significant'.

has a positive coefficient, meaning that bigger the difference in token frequency between two syllables which are also words of Kaqchikel, the easier it is to discriminate them. However, unlike the other predictors in this final model, the effect of WORD FREQUENCY is only marginally significant ($p = .084$), consistent with our overall finding that word-level factors do not have much of an effect on discrimination accuracy in our study. We believe that this effect of word frequency, though marginal, reflects the general importance of this factor in psycholinguistic processing: word frequency is consistently the strongest word-level factor in lexical retrieval tasks (such as lexical decision tasks) in a wide range of languages (Keuleers et al. 2010, 2012, Ferrand et al. 2010, Sze et al. 2014).

While all remaining fixed effects are statistically important according to our model selection procedure, differences in the size of the coefficients suggest that these predictors differ in their relative strength. STIMULUS SIMILARITY was the most important predictor ($|\beta| = 1.0720$), followed by DISTRIBUTIONAL OVERLAP ($|\beta| = 0.6320$), FUNCTIONAL LOAD ($|\beta| = 0.4653$), CATEGORY SIMILARITY (MEAN) ($|\beta| = 0.3876$) and WORD FREQUENCY ($|\beta| = 0.1848$).[13]

[13]Wedel et al. (2013) found that raw minimal pair counts (as a measure of functional load) were a better predictor of diachronic patterns of phoneme merger than lexical Δ-entropy, the metric employed here. Our implementation of functional load (systemic Δ-entropy) is proportional, but not identical, to the number of minimal pairs distinguished by a phoneme pair.

There is no current consensus as to which functional load metric is best, or whether a single metric is appropriate for analyzing all kinds of data. Given this uncertainty, we refitted our best model (3) using raw minimal pair counts and frequency-weighted minimal pair counts (as in Wedel et al. 2013) instead of lexical Δ-entropy. Functional load was not a significant predictor of consonant confusions under either of these alternative formulations (raw minimal pair counts: $z = 0.268$, $p > 0.788$; frequency-weighted minimal pair counts: $z = 0.323$, $p > 0.746$). To evaluate these two models against the model using lexical Δ-entropy, we applied two model comparison metrics, AIC and BIC: these indicated that lexical Δ-entropy provides the best fit for our data (Δ-entropy: AIC = 2406, BIC = 2452; raw minimal pair counts: AIC = 2414, BIC = 2460; frequency-weighted minimal pair counts: AIC = 2414, BIC = 2460).

7 Interim discussion

The statistical analysis in section 6 established that both STIMULUS SIMILARITY and CATEGORY SIMILARITY had an effect on discriminability in our perception study. The effect of STIMULUS SIMILARITY is unsurprising—stimuli that were acoustically more similar were, expectedly, harder to discriminate. The effect of CATEGORY SIMILARITY requires additional interpretation.

Recall that CATEGORY SIMILARITY was computed on the basis of acoustic similarity between stop categories as they occur in our corpus of spontaneous spoken Kaqchikel (sections 4.1, 5.2). We believe that this corpus provides a good approximation of the acoustic properties of Kaqchikel stops as they occur in actual, fluent speech. As such, we take the significant effect of CATEGORY SIMILARITY as an indication that phonemic categories which are acoustically well-separated in regular Kaqchikel speech are easier to discriminate in perceptual tasks.

At the theoretical level, this finding suggests that consonant discrimination is mediated by some representation of prior phonetic experience. In particular, these results are consistent with the view that speakers possess mental representations for phonemic categories which include rich phonetic detail, including (at least) some information about the acoustic properties which are typically associated with actual productions of each phoneme category in everyday speech. This claim accords with exemplar models of lexical representation, which assume that linguistic units (words, phonemes, etc.) are represented as clouds of episodic memories, which store phonetic representations of specific instances on which those units were encountered in speech (e.g. Goldinger 1996, 1998, Pierrehumbert 2001, K. Johnson 2005, Gahl & Yu 2006 and references there). Our results are also consistent with the alternative view that phonemic categories are represented in a more abstract, parametric fashion, as vectors of values along specific dimensions (e.g. VOT, closure duration, etc.) which are specified separately for each phonemic category (see Smits et al. 2006, Ernestus 2014, Pierrehumbert 2016 for discussion).

We also found that two predictors related to the lexical structure of Kaqchikel—FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP—made a significant contribution to predicting consonant confusions in our study. Notably, both of these factors are segment-level factors: the word-level factors considered here had essentially no effect on stop consonant confusions. Following Hall (2012) and others, we take FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP to be expressions of a *gradient* notion of segment-level phonemic contrast. In this sense, the relative predictability of two phonemes across contexts, and the precise number of words distinguished by those phonemes, provide a scalar characterization of how contrastive those phonemes are (i.e. how much lexical 'work' is done by the contrast between those phonemes). Our results suggest that contrasts which have a higher functional importance in Kaqchikel are also easier to discriminate, as indicated by the positive correlation between accuracy and functional load. This finding is consistent with the view that language-specific phonemic contrasts 'warp' the perceptual space in both categorical and gradient ways (e.g. Harnsberger 2000, 2001, Kataoka & Johnson 2007, Boomershine et al. 2008, Hall et al. 2014, Hall & Hume submitted and references there).

This interpretation of the results is nonetheless complicated by the finding that distributional overlap is *negatively* correlated with accuracy in our study. Such a correlation indicates

that phonemes which occur in more shared environments—that is, phonemes which are less predictable from context, and therefore *more* contrastive—are harder to discriminate. This effect is contrary to our finding for functional load, which suggests that segments which distinguish many word forms (and which are therefore *not* predictable from context) are easier to discriminate; it is also contrary to previous findings by Hall et al. (2014) and Hall & Hume (submitted) on the effect of distributional overlap on segment discrimination.

We are uncertain as to the source of this discrepancy. We first considered whether the negative correlation could be driven by the behavior of the alveolar ejective /tˀ/ alone. This segment has very low type and token frequencies in our corpora and in Kaqchikel more generally (section 4.1.3), meaning that it should have a low degree of distributional overlap with other phones. Ejective /tˀ/ is nonetheless highly perceptible—most d′ values for comparisons involving /tˀ/ are above 2, compared to a grand average of about 1.7 for all pairwise comparisons—and so this segment alone might be driving the negative correlation between accuracy and distributional overlap. However, this is not the case: when we re-run our analyses with comparisons involving /tˀ/ excluded, the effect size of DISTRIBUTIONAL OVERLAP weakens, but the negative sign does not change (β = -0.247, $p < .05$).

Alternatively, this divergent result may reflect the methods used to calculate distributional overlap in our study. As emphasized by Hall (2012), measures of distributional overlap and contextual predictability are highly sensitive to the definition of 'context' used. For example, Kaqchikel has a process which devoices syllable-final /l/ to [l̥] (e.g. *loq'ob'äl* /loqˀoɓəl/ → [loqˀoɓəl̥] 'blessing'). These two sounds are distributed completely predictably, but only if 'context' can refer to right-hand environments [__X] (i.e. [l̥] is always followed by a syllable boundary, and [l] never is). If, instead, 'context' refers only to the left-hand environment [X__], these sounds would appear to be at least partially contrastive and unpredictable (e.g. both can be preceded by [a], *wach'alal* [watʃˀalal̥] 'my family').

Previous work on this topic has computed distributional overlap using highly-specific, pre-defined contexts which either (i) reflect the structure of experimental stimuli used in the study (e.g. [a__a] in Hall et al. 2014), or (ii) reflect prior observations about the phonotactic contexts responsible for conditioning the distribution of sounds in the language under investigation (e.g. $[\_\_z]_\sigma$ in French, Hall & Hume submitted). Here, we used an inductive method (Jeffrey's divergence) defined over all possible trigram windows to compute contextual predictability (section 5.3.1). This methodological difference alone may have contributed substantially to the difference between our results and the results of Hall et al. (2014), Hall & Hume (submitted). With this in mind, we explored several other methods of computing contextual predictability, using bigram windows ([__X], [X__]) instead of trigram windows; using type rather than token frequencies (as in Hall et al. 2014, Hall & Hume submitted); and including or excluding comparisons involving /tˀ/. No combination of these methods yielded the expected positive correlation between distributional overlap and discriminability: in each case, the correlation was either non-significant or remained negative in sign.

These practical considerations aside, there is at least one other way to interpret the negative correlation between DISTRIBUTIONAL OVERLAP (degree of contrastiveness) and discriminability in our study, which again relies on exemplar dynamics. Two sounds which tend to occur in the same contexts may have more opportunities to be confused, particularly if word misperception is sensitive to statistical properties of the lexicon, including phonotactic

well-formedness (see Tang 2015). If 'confusing' sound A for sound B means erroneously storing a token of A as a token of B in the exemplar space, then frequent confusions between A and B should have the effect of making the exemplar clouds for A and B more similar over time (e.g. Wedel 2004; see also Ohala 1993). In this way, increased distributional overlap between two sounds could indirectly lead to greater confusability between those sounds by increasing the amount of overlap between their associated exemplar clouds.[14] Choosing between these possible interpretations of the effect of DISTRIBUTIONAL OVERLAP remains an open question for future research.

To reiterate, the finding that high contextual predictability (low degree of contrast) leads to greater discriminability conflicts with both our theoretical expectations (section 5.3.1) and past results on this question (Hall et al. 2014, Hall & Hume submitted). We are unsure how to interpret this result, though we note that the existence of a positive correlation between contrastiveness and discriminability has not yet been conclusively established: such a result is reported by Hall et al. (2014), Hall & Hume (submitted), but Hall (2009) finds no meaningful correlation at all between contrastiveness and discriminability (though Hall also discusses some potential issues which may have led to this null result).

In any case it seems clear, particularly for functional load, that gradient contrast has an effect on speech perception. However, the precise mechanism(s) behind patterns of contrast-driven perceptual warping remain somewhat obscure (see again Kataoka & Johnson 2007). In the following section we test two hypotheses which attempt to provide an explicit link between consonant discrimination and segment-level distributional measures (FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP) in Kaqchikel.

The first hypothesis is that functional load and distributional overlap are computed online in speech perception tasks such as ours, and that these computations can affect real-time speech processing. We do not think that this hypothesis is likely to be correct: speech perception is rapid and automatic, while the computation of functional load and distributional overlap should require substantial processing time, even if computed over some subset of the lexicon (see e.g. McClelland & Elman 1986, McClelland et al. 1986, 2006, Norris et al. 2000, Kingston et al. 2016 and references there for discussion). We nonetheless believe that this hypothesis is worthy of some consideration.

Our second hypothesis is that functional load and distributional overlap condition speech perception by shaping low-level perceptual tuning during development. By 'perceptual tuning', we refer to the fact that listeners selectively attend to those phonetic dimensions which are informative and reliable for the discrimination of phonemic categories in their native language (Holt & Lotto 2006, McGuire 2007, Davidson et al. 2007).

In the following section we attempt to disentangle these two hypotheses by investigating the timecourse of segment-level distributional factors (FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP) in our study.

[14] A question that arises is why this potential effect of similarity in the exemplar space isn't already captured by our measure of CATEGORY SIMILARITY, under the assumption that similarity in an acoustic corpus is a good reflection of similarity between abstract exemplar clouds (e.g. Pierrehumbert 2001). One possibility is that distributional overlap affects the shape of exemplar distributions in a manner which is not captured by our relatively simple measure of category similarity (distance between centroids). Investigating this issue in detail would take us too far beyond the goals of the present article.

8 The timecourse of experience-based effects

Speech perception can be decomposed into at least three distinct tasks: auditory/acoustic processing; phone-level processing; and lexical retrieval (e.g. Pisoni & Tash 1974, Pisoni 1975, Fox 1984, Pitt & Samuel 1993, Werker & Logan 1985, Babel & Johnson 2010 and references there). The first of these tasks, auditory/acoustic processing, involves mechanisms which are basically physiological in nature. As such, this aspect of speech processing is not expected to be substantially affected by the listener's native language. Phone-level processing (sometimes called 'phonetic' or 'phonemic' processing, e.g. Pisoni 1973, Werker & Tees 1984, Werker & Logan 1985) involves the categorization of speech sounds into appropriate phonemes and/or allophones. This type of processing differs from auditory/acoustic processing in that it is necessarily conditioned by the listener's native language, and is therefore expected to show sensitivity to past linguistic experience. Such sensitivity is also expected for any aspect of speech processing that involves lexical access, as languages (and speakers) obviously differ in their vocabularies.

Researchers disagree as to the relative independence of each of these aspects of speech processing (see McClelland & Elman 1986, McClelland et al. 1986, 2006, Norris et al. 2000, Kingston 2005, Kingston et al. 2016 for discussion and further references). There is nonetheless a broad consensus that native-language influences on speech perception emerge relatively late in the timecourse of speech processing. This is particularly true of lexical effects, which tend to influence speech perception sometime after the initiation of phone-level processing (Fox 1984; though cf. Kingston et al. 2016 and work cited there).

Assuming that these stages of speech processing have the rough temporal sequencing suggested by prior work (acoustic/auditory ⇒ phone-level ⇒ lexical), we can at least tentatively diagnose the mechanism behind the segment-level statistical effects in our study (functional load and distributional overlap) by investigating when in the course of speech processing those effects arise. If functional load (FL) and distributional overlap (DO) are computed online during speech perception, through some process of lexical access or lexical sampling, then the effect of these predictors should emerge relatively late. We would then expect stronger effects of FL/DO at slower response times. If, on the other hand, FL/DO affect speech processing by shaping low-level perceptual tuning (e.g. cue weighting) during acquisition, then we might expect to see the influence of these predictors even at relatively fast response times.

8.1 Predictions

Among the significant predictors in our final model (3), STIMULUS SIMILARITY corresponds most closely to the kind of information that would be processed during the acoustic/auditory stage of speech perception. We thus expect that STIMULUS SIMILARITY should have a robust effect on response accuracy at even the fastest response times. CATEGORY SIMILARITY, a measure which refers to language-specific phonetic distributions associated with individual phoneme categories, should emerge no earlier than than the purely acoustic measure of STIMULUS SIMILARITY. If functional load and distributional overlap are computed online, they should begin to affect responses at a later stage than either STIMULUS SIMILARITY or CATEGORY SIMILARITY.

Apart from the relative onset of these effects, we might also find differences in how the influence of each factor changes over time. Even if all of the factors in the model begin to affect response accuracy at about the same point, some factors might still grow in strength over time, while others weaken instead. In particular, factors involving lexical access might be more evident at longer response latencies, under the assumption that the strength of lexical activation gradually increases over time, such that lexical factors influence phone-level activation more strongly at later stages of processing (e.g. Kingston et al. 2016).

8.2 Statistical modeling

To carry out a timecourse analysis of our results we fit a new regression model based on our previous best model (3). Five interaction terms were added to test whether the significant fixed effects in (3) interact with response time in predicting participant accuracy. RESPONSE TIME was also added as a fixed effect, consistent with the standard practice of including simple effects for any predictor included in an interaction term. Nested model comparison shows that model fit is significantly improved when these five interaction terms are included (Table 3; $\chi^2(5) = 23.6$, $p < .001$).

We also performed a separate timecourse analysis treating RESPONSE TIME as a discrete variable rather than a continuous one. Dichotomizing continuous variables is often discouraged (Baayen 2008: 259), but we performed this additional analysis because it more closely resembles the treatment of timecourse effects on speech processing in some previous work (Kingston 2005, Babel & Johnson 2010). First, each participant's responses were divided into three equally-sized bins (i.e. by-participant terciles): fast responses (EARLY), medium-speed responses (MIDDLE), and slow responses (LATE). The mean response times for each bin (across participants) were about 400ms, 650ms, and 1200ms; the first bin falls roughly in the range of response times associated with auditory processing, while the latter two fall in the range associated with phone-level and/or lexical processing (e.g. Fox 1984, Werker & Tees 1984, Werker & Logan 1985, Babel & Johnson 2010). We Helmert-coded this discrete, three-level timecourse predictor, and re-fit the model (3) using the same structure used for our continuous response time predictor. This analysis yielded the same qualitative results as the analysis which treated timecourse as a continuous predictor. We report only the continuous model below.[15]

The significant interactions reported in Table 3 suggest that the strength of our acoustic and distributional predictors did vary as a function of response time. To dig deeper into the interaction between these factors and response time, we fit a separate regression model for each of the response time tercile bins described above (Table 4), using the same model structure (3) which we used in the analysis of the full data set.

8.3 Statistical results

Table 3 presents the regression statistics of the model with interaction terms between the five fixed effects and RESPONSE TIME. We first note that WORD FREQUENCY and its interaction

[15]The model treating response time as a continuous variable has a significantly better fit to our data than the model which uses discrete response time bins (continuous response time model: AIC = 2391, BIC = 2466; discrete response time model: AIC = 2399, BIC = 2479).

term with RESPONSE TIME do not reach statistical significance. The other four interaction terms do reach statistical significance. Crucially, the coefficients of these four significant interaction terms indicate that the four fixed effects decrease in strength as response times increase.

Table 4 presents the regression statistics for each of the three regression models, grouped by response time bin. We first note that WORD FREQUENCY does not reach statistical significance in any response time bin. The other four predictors—STIMULUS SIMILARITY, CATEGORY SIMILARITY, FUNCTIONAL LOAD, and DISTRIBUTIONAL OVERLAP—have consistent effects across all three response time bins. Each of these four predictors influence response accuracy even in the earliest bin. Additionally, all of these predictors (including the insignificant predictor WORD FREQUENCY) decrease in strength as response times increase. This decrease in strength is evident in both the magnitude of the effects (the values of the β coefficients) and the level of statistical significance reached. Together these findings suggest that all four significant factors kick in early, but decrease in strength over time.

	β	SE(β)	$\lvert z \rvert$	p-value
(Intercept)	0.8248	0.1635	5.044	$< .001^{*}$
STIMULUS SIMILARITY	-1.0705	0.1165	9.193	$< .001^{***}$
CATEGORY SIMILARITY (MEAN)	-0.3975	0.1257	3.163	$.002^{**}$
FUNCTIONAL LOAD	0.4820	0.1668	2.889	$.004^{**}$
DISTRIBUTIONAL OVERLAP	-0.6544	0.1630	4.015	$< .001^{***}$
WORD FREQUENCY (ABS. DIFF.)	0.1705	0.1084	1.573	$.116^{n.s.}$
RESPONSE TIME	-0.1437	0.0637	2.254	$.024^{*}$
STIMULUS SIMILARITY:RESPONSE TIME	0.1877	0.0742	2.528	$.011^{*}$
CATEGORY SIMILARITY (MEAN):RESPONSE TIME	0.1591	0.0794	2.005	$.045^{*}$
FUNCTIONAL LOAD:RESPONSE TIME	-0.2737	0.1063	2.575	$.010^{*}$
DISTRIBUTIONAL OVERLAP:RESPONSE TIME	0.3389	0.1076	3.149	$.002^{**}$
WORD FREQUENCY (ABS. DIFF.):RESPONSE TIME	0.0132	0.0686	0.193	$.8465^{n.s.}$

Table 3: Regression statistics of the best model (3) with added interaction terms between the five fixed effects and by-participant response time predicting response accuracy (correct: 1, incorrect: 0) in log odds space.

8.4 Interpretation of timecourse analysis

This timecourse analysis shows that three experience-based factors (CATEGORY SIMILARITY, FUNCTIONAL LOAD, and DISTRIBUTIONAL OVERLAP) began to affect discrimination at about

	Early ($\mu \approx$ 400ms)	Middle ($\mu \approx$ 650ms)	Late ($\mu \approx$ 1200ms)
	β	β	β
stimulus similarity	-1.4515***	-1.1651***	-0.7465***
category similarity	-0.6544**	-0.3020.	-0.2876*
functional load	0.9001**	0.4116.	0.2853.
distributional overlap	-1.1437***	-0.8765***	-0.2797.
word frequency (abs. diff.)	$0.2671^{n.s.}$	$0.2314^{n.s.}$	$0.0607^{n.s.}$

Table 4: Regression statistics for the fixed effects of three regression models, computed over by-participant response time terciles, predicting response accuracy (correct: 1, incorrect: 0) in log odds space.

the same early timepoint as the acoustic/auditory factor stimulus similarity.

For present purposes, the most important result is that functional load and distributional overlap influenced response accuracy even at very fast response times (those in the early tercile bin). This result is consistent with the view that these two 'lexical' measures impinge on speech perception somewhat indirectly, most likely by influencing which acoustic dimensions speakers attend more closely to during phone-level processing. If functional load and distributional overlap condition speech perception through perceptual tuning, these factors are *expected* to show the same timecourse as category similarity, as all three measures reflect perceptual processes which in some way refer to the phonetic dimensions that distinguish phoneme categories in the listener's native language.

We cannot completely rule out the possibility that functional load and distributional overlap are computed online during speech perception. For one, the inter-stimulus interval in this study (up to 800ms) may have been sufficiently long that listeners were able to carry out some form of lexical access even for fairly quick responses. Nevertheless, we believe that this interpretation of our results is at odds with several observations. First, the AX discrimination task used in this study neither required nor encouraged lexical access, particularly because most stimuli were not actual words of Kaqchikel. Second, we think it is inherently unlikely that listeners carry out the large-scale lexical access that would be needed to accurately compute measures like functional load online. Speech processing is simply too rapid to involve lexical access at this scale during real-time listening. This argument is bolstered by the observation that the effect of these 'lexical' measures emerged early and *weakened* over time; were some form of bulk lexical access involved, we should expect to see the strength of these measures increase over time instead, as more of the lexicon is accessed and analyzed.

9 Discussion

9.1 Theoretical contributions

The core theoretical contributions of this article are twofold. First, we have demonstrated experimentally that prior linguistic experience affects speech perception, not simply because different languages have different phonemic inventories (e.g. Werker & Tees 1984, Werker & Logan 1985), but also because languages differ in the fine phonetic details associated with phonemic categories, as well as in their lexical structure and patterns of usage. These results replicate and extend past research showing that highly specific statistical patterns in a listener's native language can have extensive effects on perceptual processing, even in experimental tasks that do not obviously require lexical access.

Importantly, our investigation has established these results in the context of a language—Kaqchikel Maya—which is sociolinguistically and structurally very different from the majority languages which are most often studied in speech perception research (section 1). We hope that this work will encourage researchers to continue expanding the speech perception literature, and the phonetics literature more generally, to include a wider range of lesser-studied languages. Only in this way can we establish cross-linguistically valid theories of speech perception and production.

9.2 Methodological contributions

In this study, we replicated several findings of experience-based effects in speech perception which have previously been demonstrated for majority languages using much richer resources (e.g. substantially larger written corpora, section 4.2.2). We take this result to be an indirect validation of the use of small corpora in speech perception studies. Despite their shortcomings, small, noisy corpora can make valuable contributions to speech perception research, provided they are carefully processed beforehand. Our results also supply a positive answer to the general question of whether reliable speech perception research can be conducted in the field, outside of highly-controlled laboratory settings (Whalen & McDonough 2015; see also DiCanio 2014 for an excellent recent example of this kind of research).

To further support this claim, we now compare our findings to a similar study conducting with a majority language (French) in a laboratory setting. Hall & Hume (submitted) investigated segment-level statistical effects on the discriminability of French vowels. They considered many of the same predictors we investigated in our study, including STIMULUS SIMILARITY, FUNCTIONAL LOAD, DISTRIBUTIONAL OVERLAP and SEGMENTAL TOKEN FREQUENCY; this parallelism allows us to compare the two studies rather directly.

With the exception of CATEGORY SIMILARITY and WORD FREQUENCY (which were not examined by Hall & Hume), the significant predictors in our study (STIMULUS SIMILARITY, FUNCTIONAL LOAD, and DISTRIBUTIONAL OVERLAP) were also statistically significant predictors of vowel discrimination in Hall & Hume (submitted) (though the direction of the effect for DISTRIBUTIONAL OVERLAP was different in the two studies). SEGMENT TOKEN FREQUENCY did not reach significance in either our study or in Hall & Hume submitted. In terms of the relative importance of these predictors, both studies found that stimulus similarity played the strongest role, followed by predictors related to phonological contrastiveness

(FUNCTIONAL LOAD and DISTRIBUTIONAL OVERLAP), with DISTRIBUTIONAL OVERLAP being a better predictor of response accuracy than FUNCTIONAL LOAD (though only when /tˀ/ is included in the analysis). We find the parallelism between these results to be rather encouraging, especially given the following differences between the two studies:

1. **Task**: Hall & Hume used a multiple forced-choice identification task, while our study used an AX discrimination task.
2. **Target segments**: Hall & Hume examined vowels, while our study examined consonants.
3. **Stimulus presentation**: Hall & Hume presented their stimuli without any masking noise, while we presented our stimuli in speech-shaped noise at a 0dB SNR.
4. **Experimental setting**: Hall & Hume tested their participants in a controlled laboratory setting in a sound-attenuated booth, while we tested our participants in a quiet room which was not sound-attenuated.
5. **Language**: Hall & Hume examined French, a Romance language, while we examined Kaqchikel, a Mayan language.
6. **Culture**: the participants in Hall & Hume's study were likely to have some experience with psychological experiments, and extensive experience with computers. Our participants did not in general have such experience.
7. **Quality of the corpus estimates**: the segment-level predictors in Hall & Hume (submitted) were estimated using a written corpus that is very large (65.1 million words), well-balanced, and highly speech-like (compiled from books, and subtitles of films and TV shows). Our segment-level predictors were estimated using a written corpus that is small (0.7 million words), unbalanced, and not particularly speech-like (mostly governmental and religious documents).

Despite these differences, which are varied and numerous, the two studies arrive at strikingly similar conclusions about the kinds of predictors which affect speech perception, as well as their relative importance. This comparison thus reaffirms our claim that conducting speech perception research in the field can result in findings that are comparable to those done in a laboratory.

However, it should also be noted that our results are substantially 'noisier' than the results of Hall & Hume (submitted). In particular, the fixed effects components in our statistical model (section 6) capture 23.2% of the variance in our data; in contrast, the fixed effects components in the statistical model reported in Hall & Hume (submitted) capture as much as 82% of the variance in their study.[16] It is unclear to us which differences between the studies (including, but not limited to those outlined above) could have led to such a dramatic difference in model fit. Answering this question would require several follow-up studies which reduce the methodological differences with Hall & Hume (submitted), a project we leave for future research.

[16]The proportion of variance captured by fixed effects in our model was computed with the function `r.squaredGLMM()`, part of the `MuMIn` library in R (Nakagawa & Schielzeth 2013, P. C. Johnson 2014, Bartoń 2014). This function returns both marginal R^2_{GLMM} and conditional R^2_{GLMM}. The reported variance is the marginal R^2_{GLMM}, which represents the variance explained by fixed factors.

Acknowledgements

First and foremost, we thank the Kaqchikel speakers who participated in the perception study described here, as well as in the development of our spoken and written corpora. Asociación Ceiba (`http://www.ceibaguate.org.gt/`) kindly provided recording space for the production of our spoken corpus, which was collected with the assistance of the Comunidad Lingüística Kaqchikel (`http://kaqchikel.almg.org.gt/`). Centro Educativo Maya Aj Sya (`https://mayaajsya.wordpress.com/about/`) gave us both space and extensive support for carrying out our perception experiment. *Janila matyöx chiwe iwonojel*! We also thank Robert Henderson for logistical help in conducting the perception experiment. We are grateful to Doug Whalen (Haskins Laboratories), Jason Shaw (Yale University) and Uriel Cohen Priva (Brown University) for detailed feedback at various stages of the development of this project. In addition, we thank audiences at the Yale Phonetics/Phonology Reading Group, Speech Science Forum at University College London, Haskins Laboratories, the University of Hong Kong, Brown University, The Hong Kong Polytechnic University, The Education University of Hong Kong, UC Santa Cruz, *Workshop on Structure and Constituency in Languages of the Americas 21*, *Sound Systems of Mexico and Central America II*, *Form and Analysis in Mayan Linguistics IV*, the *91st Annual meeting of the Linguistic Society of America*, and the *24th Manchester Phonology Meeting*.

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 et al. 2010: 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 et al. 2000: 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.

Appendix B: model construction and selection

Before the initial model (section 6) was assessed, we first evaluated whether collinearity of all the fixed effects would be an issue. Using the function `collin.fnc()` in the `languageR`

library (Baayen 2013), the condition number (Belsley et al. 2005) was calculated for all the fixed effects, and a value of 4.567 was obtained. Given that the condition number is between 0 and 6, there is very little collinearity between our fixed effects (Baayen 2008: 198-200). Therefore, we can be confident that each of the fixed effects is a relatively independent component in the regression model.

In constructing our initial model, we did not follow Barr et al.'s (2013) recommendation to fit the most complex, convergent random effect structure for our model. This practice has been critiqued elsewhere because it can lead to uninterpretable models (Baayen et al. 2017) with reduced statistical power (Matuschek et al. 2017). Instead, the random effect structure of the initial superset model was determined by the authors on the basis of the empirical and theoretical considerations discussed in section 6. The superset model contains all fixed effects, without interaction terms between them, as well as the random intercepts and slopes mentioned above.

(4) Superset model syntax in LME4

> Accuracy ~ Stimulus Similarity + Category Similarity (Mean) + Category Similarity (SD) + Segmental Frequency + Functional Load + Distributional Overlap + Wordhood + Word Frequency + Neighborhood Density + Average Neighborhood Frequency + Bigram frequency + response time + (1 | Unordered Stimulus Pair) + (1 | List) + (1 | Stimulus Order) + (1 |Onset-Coda) + (1 + Segmental Frequency + Functional Load + Segment Environment Distribution + Wordhood + Word Frequency + Neighborhood Density + Average Neighborhood Frequency + Bigram frequency | Participant)

This initial model was then simplified following a step-down, data-driven model selection procedure which compared nested models using the backward best-path algorithm (e.g. Gorman & Johnson 2013, Barr et al. 2013). We began by simplifying the random effects structure of the model. First, we generated a set of models which were minimally simpler than the superset model, differing only in the ommission of one of the random effects. Each of these models was then compared to the superset model using a likelihood ratio test, computed with the `anova()` function. If the likelihood ratio test resulted in a p-value of 0.1 or higher, the simpler model was taken to be an improvement on the superset model.

We chose a relatively liberal threshold of $\alpha = 0.1$ to be conservative in our model selection procedure, preferring to include potentially relevant predictors in the final model if they were reasonably well-justified. In the case that there were multiple subset models which exceeded this α threshold in comparison with the superset model, the subset model with the strongest evidence (the highest p-value) was selected. The random intercepts for both PARTICIPANT and item (UNORDERED STIMULUS PAIR) were never considered for exclusion, as it is standard practice to include these random effects in models of this type (e.g. Jaeger 2008). This procedure was then repeated, with successive simplification of the random effects structure, until no subset model exceeded the α threshold of 0.1 in comparison with the immediate superset model.

After the best random effect structure was determined, the same steps were repeated to determine the best fixed effects structure. This procedure alternated between random and

fixed effect structures until the model could not be reduced any further. The resultant, 'best' model is given in (5).

(5) Best model

> Accuracy ~ Stimulus Similarity + Category Similarity (Mean) + Functional Load + Distributional Overlap + Word Frequency + (1 | Unordered Stimulus Pair) + (1 | Participant)

The overall fit of the model, R^2_{GLMM}, was calculated using the `r.squaredGLMM()` function in the `MuMIn` library (Nakagawa & Schielzeth 2013, P. C. Johnson 2014, Bartoń 2014). The marginal R^2_{GLMM} is 23.2% and conditional R^2_{GLMM} is 49.9%. Marginal R^2_{GLMM} represents the variance explained by fixed factors and conditional R^2_{GLMM} represents the variance explained by both fixed and random factors. A sizeable portion of the variance (23.2%) was explained with only five fixed factors, suggesting that our relatively simple model is unlikely to be over-fitting the data.

Appendix C: Non-significant predictors in the AX discrimination study

Non-significant segment-level predictors

Segmental frequency

Segmental frequency is simply the number of occurrences of a particular segment type in some representative corpus. Segmental frequency has been shown to influence the location of the boundary between phonemic categories in a discrimination task (Kataoka & Johnson 2007). Crucially for this study, the difference in segmental frequency between two phonemes may also condition discrimination accuracy above and beyond the acoustic differences between those two phonemes (Bundgaard-Nielsen & Baker 2014, Bundgaard-Nielsen et al. 2015).

Further, segmental frequency has been shown to influence the probability of a segment being misperceived in English, as well as the quality of the segment that listeners incorrectly perceive when making a perceptual error. Using naturalistic misperception data in English, Tang (2015: Ch. 4) found that (i) more frequent segments are more likely to be misperceived, and (ii) when a segment is misperceived, listeners are more likely to report hearing a relatively frequent segment.

In our case, segment frequency is calculated over each stop consonant type regardless of the vowel that follows it and its position in a syllable (onset or coda).

Non-significant word-level predictors

Neighborhood density and neighborhood frequency

The phonological neighbors of a given word w are usually defined as words which differ from w only by the deletion, addition, or substitution of one phoneme. The phonological neighborhood density of a word w is simply defined as the number of neighbors w has.

It has been demonstrated that frequency-weighted lexical neighborhood density (the number of neighbors of a given word, weighted by their token frequencies) can affect the phonemic categorization of a stimulus much like the lexical status of a stimulus can (Ganong 1980, Newman et al. 1997, 2005).

More generally, NEIGHBORHOOD DENSITY and AVERAGE NEIGHBORHOOD FREQUENCY (the average of the log token frequencies of an item's neighbours) can have an inhibitory effect on spoken/visual word recognition, such that words with high neighborhood density and/or high average neighborhood frequency are more difficult to recognize (Luce 1986, Luce & Pisoni 1998, Grainger & Segui 1990). Furthermore, when a word is recognized incorrectly, the neighborhood density of the perceived word is typically similar to the neighborhood density of the intended word (Vitevitch 2002). Consequently, even if a stimulus in our study were incorrectly perceived, its neighborhood density could still bias the participants' responses.

In order to keep our results directly comparable to a large body of past work in speech perception, we focused on just neighborhood density and average neighborhood frequency—two common measures of neighborhood structure—as potential neighborhood-related predictors of stimulus confusability. For general discussion, as well as other techniques for calculating neighborhood density and related measures, see Luce (1986), Bailey & Hahn (2001), Yarkoni et al. (2008), Yao (2011), Gahl & Strand (2016), Vitevitch & Luce (2016).

Bigram frequency

BIGRAM FREQUENCY refers to the frequency of occurrence of each two-phoneme sequence (bigram) in some corpus. In the context of our study, the bigram frequency of a stimulus can be interpreted as the number of times a given stop consonant is preceded/followed by a given vowel in our written corpus. Our BIGRAM FREQUENCY predictor characterized the difference in bigram frequency between the two stimuli presented on a given trial.

Bigram frequency is known to play a role in non-word acceptability tasks (Albright 2009) as well as in the recognition of both real words and non-words (Rice & Robinson 1975, Vitevitch & Luce 1999). In terms of our stimuli, bigram frequency can be interpreted as an approximate estimate of overall word-likeness.

Given that our stimuli consisted of monosyllables, the bigram frequency of a given stimulus is likely to correlate with an estimate of syllable frequency. It is well-known that syllable frequency influences word recognition, with an inhibitory effect: words with high-frequency initial syllables are recognized more slowly than words with low-frequency initial syllables. This effect can be understood in a cohort model of lexical activation: a high-frequency initial syllable should activate a larger number of competing lexical candidates, thus slowing word recognition. This effect is found in both visual (Carreiras et al. 1993, Barber et al. 2004) and spoken word recognition (González-Alvarez & Palomar-García 2016).

Appendix D: d′ scores for all target stop contrasts

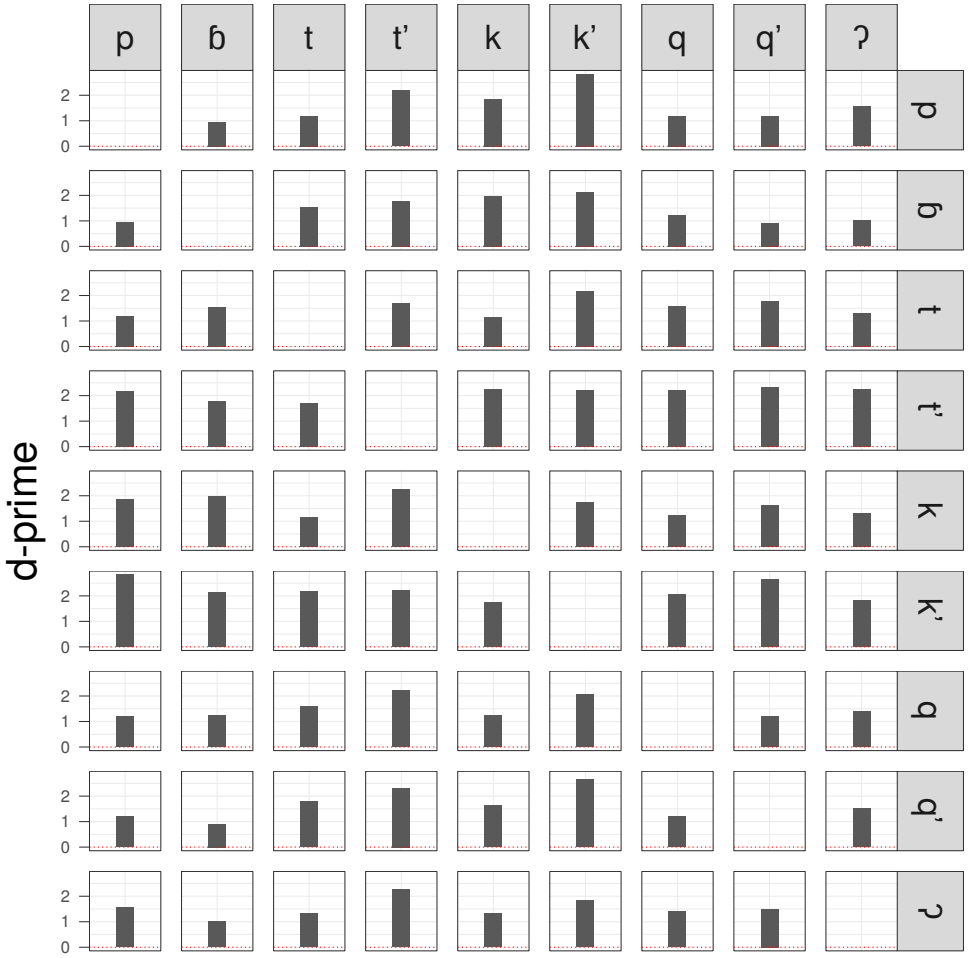

Figure 4: Plot of d′ scores for all target stop contrasts in the AX discrimination study, collapsed across vowel context and syllable position

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(01). 9–41.

Atkins, S., J. Clear & N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1–16.

Baayen, R. Harald. 2008. *Analyzing linguistic data: a practical introduction to statistics using r*. Cambridge University Press.

Baayen, R. Harald. 2013. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. R package version 1.4.1. `https://CRAN.R-project.org/package=languageR`.

Baayen, R. Harald, Shravan Vasishth, Reinhold Kliegl & Douglas Bates. 2017. The cave of shadows: addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234.

Babel, Molly & Keith Johnson. 2010. Accessing psycho-acoustic perception and language-specific perception with speech sounds. *Laboratory phonology* 1(1). 179–205.

Baese-Berk, Melissa & Matthew Goldrick. 2009. Mechanisms of interaction in speech production. *Language and cognitive processes* 24(4). 527–554.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4). 568–591.

Barber, Horacio, Marta Vergara & Manuel Carreiras. 2004. Syllable-frequency effects in visual word recognition: evidence from erps. *Neuroreport* 15(3). 545–548.

Barr, Dale, Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 68(3). 255–278. `https://doi.org/http://dx.doi.org/10.1016/j.jml.2012.11.001.//www.sciencedirect.com/science/article/pii/S0749596X12001180`.

Barrett, Rusty. 1999. *A grammar of Sipakapense Maya*. University of Texas at Austin dissertation.

Bartoń, Kamil. 2014. *MuMIn: Multi-model inference*. R package version 1.10.0. `http://CRAN.R-project.org/package=MuMIn`.

Bates, Douglas, Martin Maechler & Ben Bolker. 2011. *lme4: Linear mixed-effects models using S4 classes*. Version 0.999375-41, retrieved from http://CRAN.R-project.org/package=lme4.

Belsley, David A., Edwin Kuh & Roy E. Welsch. 2005. Detecting and Assessing Collinearity. In *Regression Diagnostics*, 85–191. John Wiley & Sons, Inc. `https://doi.org/10.1002/0471725153.ch3`. `http://dx.doi.org/10.1002/0471725153.ch3`.

Benkí, José R. 2003. Analysis of English nonsense syllable recognition in noise. *Phonetica* 60(2). 129–157. `https://doi.org/10.1159/000071450`.

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

Bennett, Ryan. 2010. Contrast and laryngeal states in Tz'utujil. In Grant McGuire (ed.), *UC Santa Cruz Linguistics Research Center annual report*, 93–120. Available online at http://people.ucsc.edu/∼gmcguir1/LabReport/BennettLRC.pdf. Santa Cruz, CA: LRC Publications.

Bennett, Ryan. 2016. Mayan phonology. *Language and Linguistics Compass* 10(10). 469–514.

Bennett, Ryan & Juan Ajsivinac Sian. In preparation, c. *Un corpus fonético del kaqchikel de Sololá, Guatemala: narrativas espontáneas*. Corpus electrónico, grabado en 2013.

Bennett, Ryan, Jessica Coon & Robert Henderson. 2016. Introduction to Mayan linguistics. *Language and Linguistics Compass* 10(10). 1–14.

Bennett, Ryan, Kevin Tang & Juan Ajsivinac Sian. Submitted. Laryngeal co-occurrence restrictions as constraints on sub-segmental articulatory structure.

Best, Catherine T. 1995. A direct realist view of cross-language speech perception. 171–204.

Best, Catherine T., Gerald McRoberts & Elizabeth Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America* 109(2). 775–794.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Bladon, Anthony. 1986. Phonetics for hearers. In Graham McGregor (ed.), *Language for hearers*, 1–24. Oxford, UK: Pergamon Press.

Boersma, Paul & David Weenink. 2016. *Praat: doing phonetics by computer (Version 6.0.23)*. Computer program. Retrieved from http://www.praat.org/.

Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume & Keith Johnson. 2008. The impact of allophony versus contrast on speech perception. In Peter Avery, B. Elan Dresher & Keren Rice (eds.), *Contrast in phonology: theory, perception, acquisition*, 145–171. Berlin: de Gruyter.

Broadbent, Donald. 1967. Word-frequency effect and response bias. *Psychological Review; Psychological Review* 74(1). 1–15.

Brody, Michal. 2004. *The fixed word, the moving tongue: variation in written Yucatec Maya and the meandering evolution toward unified norms*. Austin, TX: University of Texas Austin dissertation.

Browman, Catherine & Louis Goldstein. 1986. Towards an articulatory phonology. *Phonology yearbook* 3(21). 219–252.

Browman, Catherine & Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology* 6(2). 201–251.

Browman, Catherine & Louis Goldstein. 1992. Articulatory phonology: an overview. *Phonetica* 49(3-4). 155–180.

Brown, Charles R. & Herbert Rubenstein. 1961. Test of response bias explanation of word-frequency effect. *Science* 133(3448). 280–281.

Brown, R. McKenna, Judith Maxwell & Walter Little. 2010. *La ütz awäch?: introduction to Kaqchikel Maya language*. Austin, TX: University of Texas Press.

Brysbaert, Marc & Kevin Diependaele. 2013. Dealing with zero word frequencies: a review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods* 45(2). 422–430.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods* 41(4). 977–990.

Bundgaard-Nielsen, Rikke L. & Brett J. Baker. 2014. Frequency in the input affects perception of phonological contrasts for native speakers. In *Proceedings of the 15th Australasian International Speech Science and Technology Conference*, 205–208.

Bundgaard-Nielsen, Rikke L., Brett J. Baker, Christian H. Kroos, Mark Harvey & Catherine T. Best. 2015. Discrimination of multiple coronal stop contrasts in Wubuy (Australia): a natural referent consonant account. *PLOS ONE* 10(12). 1–30. `https://doi.org/10.1371/journal.pone.0142054`. `http://dx.doi.org/10.1371%2Fjournal.pone.0142054`.

Calamaro, Shira & Gaja Jarosz. 2015. Learning general phonological rules from distributional information: a computational model. *Cognitive science* 39(3). 647–666.

Campbell, Lyle. 1977. *Quichean linguistic prehistory*. Vol. 81 (University of California Publications in Linguistics). Berkeley, CA: University of California Press.

Carreiras, Manuel, Carlos J. Alvarez & Manuel de Vega. 1993. Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language* 32(6). 766–780. `https://doi.org/http://dx.doi.org/10.1006/jmla.1993.1038`. `http://www.sciencedirect.com/science/article/pii/S0749596X83710387`.

Chacach Cutzal, Martín. 1990. Una descripción fonológica y morfológica del kaqchikel. In Nora England & Stephen Elliott (eds.), *Lecturas sobre la lingüística maya*, 145–190. Antigua, Guatemala: Centro de Investigaciones Regionales de Mesoamérica.

Chang, Steve, Madelaine Plauché & John Ohala. 2001. Markedness and consonant confusion asymmetries. In Keith Johnson & Elizabeth Hume (eds.), *The role of speech perception in phonology*, 79–101. New York: Academic Press.

Cojtí Macario, Narciso & Margarita Lopez. 1990. Variación dialectal del idioma kaqchikel. In Nora England & Stephen Elliott (eds.), *Lecturas sobre la lingüística maya*, 193–220. Antigua, Guatemala: Centro de Investigaciones Regionales de Mesoamérica.

Coon, Jessica. 2016. Mayan morphosyntax. *Language and Linguistics Compass* 10(10). 515–550.

Cowan, Nelson & Philip A Morse. 1986. The use of auditory and phonetic memory in vowel discrimination. *The Journal of the Acoustical Society of America* 79(2). 500–507.

Cutler, Anne. 2012. *Native listening: language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

Cutler, Anne, Andrea Weber, Roel Smits & Nicole Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116(6). 3668–3678.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 2009. *TiMBL: Tilburg Memory-Based Learner*. Reference Guide Version 6.2. ILK Technical Report – ILK 09-01.

Dar, Mariam, Tamar Keren-Portnoy & Marilyn Vihman. 2018. An order effect in English infants' discrimination of an Urdu affricate contrast. *Journal of Phonetics* 67. 49–64.

Davidson, D.J. & Andrea E. Martin. 2013. Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta psychologica* 144(1). 83–96.

Davidson, Lisa, Jason Shaw & Tuuli Adams. 2007. The effect of word learning on the perception of non-native consonant sequences. *The Journal of the Acoustical Society of America* 122(6). 3697–3709. `https://doi.org/10.1121/1.2801548`. `http://dx.doi.org/10.1121/1.2801548`.

DiCanio, Christian. 2014. Cue weight in the perception of Trique glottal consonants. *Journal of the Acoustical Society of America* 135(2). 884–895.

DiCanio, Christian, Hosung Nam, D.H. Whalen, H. Timothy Bunnell, Jonathan D. Amith & Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134(3). 2235–2246.

Dockum, Rikker & Ethan Campbell-Taylor. 2017. *Minimum sufficient wordlist size for phonological typology*. Ms., Yale University.

Dubno, Judy & Harry Levitt. 1981. Predicting consonant confusions from acoustic analysis. *The Journal of the Acoustical Society of America* 69(1). 249–261.

DuBois, John W. 1981. *The Sacapultec language*. University of California, Berkeley dissertation.

Dunbar, Ewan & William J. Idsardi. 2010. Review of Daniel Silverman (2006). A critical introduction to phonology: of sound, mind, and body. London & New York: Continuum. Pp. xii+260. *Phonology* 27(2). 325–331.

El Hattab, Hakim. 2016. *reveal.js*. `https://github.com/hakimel/reveal.js/`.

England, Nora. 1983. *A grammar of Mam, a Mayan language*. Austin, Texas: University of Texas Press.

England, Nora. 1996. The role of language standardization in revitalization. In Edward Fischer & R. McKenna Brown (eds.), *Maya cultural activism in Guatemala*, 178–194. Austin, TX: University of Texas Press.

England, Nora. 2001. *Introducción a la gramática de los idiomas mayas*. Ciudad de Guatemala, Guatemala: Cholsamaj.

England, Nora. 2003. Mayan language revival and revitalization politics: linguists and linguistic ideologies. *American Anthropologist* 105(4). 733–743.

Ernestus, Mirjam. 2014. Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua* 142. 27–41.

Felty, Robert Albert, Adam Buchwald, Thomas Gruenenfelder & David Pisoni. 2013. Misperceptions of spoken words: data from a random sample of American English words. *The Journal of the Acoustical Society of America* 134(1). 572–585.

Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova & Christophe Pallier. 2010. The french lexicon project: lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods* 42(2). 488–496.

Fischer, Edward & R. McKenna Brown (eds.). 1996. *Maya cultural activism in Guatemala*. Austin, TX: University of Texas Press.

Fox, Robert. 1984. Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human perception and performance* 10(4). 526–540.

Fre Woldu, Kiros. 1985. *The perception and production of Tigrinya stops*. Vol. 13 (Reports from Uppsala University Department of Linguistics). Uppsala: Department of Linguistics, Uppsala University.

Fujimura, Osamu, Marian Macchi & Lynn Streeter. 1978. Perception of stop consonants with conflicting transitional cues. *Language and Speech* 21. 337–343.

Gahl, Susanne & Julia F. Strand. 2016. Many neighborhoods: phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* 89. 162–178.

Gahl, Susanne & Alan C.L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review* 23(3). 213–216.

Gallagher, Gillian. 2010a. Perceptual distinctness and long-distance laryngeal restrictions. *Phonology* 27(3). 435–480.

Gallagher, Gillian. 2010b. *The perceptual basis of long-distance laryngeal restrictions*. Massachusetts Institute of Technology dissertation.

Gallagher, Gillian. 2011. Acoustic and articulatory features in phonology–the case for [long VOT]. *The Linguistic Review* 28(3). 281–313.

Gallagher, Gillian. 2012. Perceptual similarity in non-local laryngeal restrictions. *Lingua* 122(2). 112–124.

Gallagher, Gillian. 2014. An identity bias in phonotactics: evidence from Cochabamba Quechua. *Laboratory Phonology* 5(3). 337–378.

Ganong, William. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6(1). 110.

García Matzar, Pedro Oscar, Valerio Toj Cotzajay & Domingo Coc Tuiz. 1999. *Gramática del idioma Kaqchikel*. Antigua, Guatemala: Proyecto Lingüístico Francisco Marroquín.

Gasser, Emily & Claire Bowern. 2014. Revisiting phonotactic generalizations in Australian languages. *Proceedings of the Annual Meetings on Phonology* 1(1). http://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/17.

Goldinger, Stephen D. 1996. Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of experimental psychology: Learning, memory, and cognition* 22(5). 1166–1183.

Goldinger, Stephen D. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review* 105(2). 251–279.

Goldrick, Matthew, Charlotte Vaughn & Amanda Murphy. 2013. The effects of lexical neighbors on stop consonant articulation. *Journal of the Acoustical Society of America* 134(2). EL172–EL177.

González-Alvarez, Julio & María-Angeles Palomar-García. 2016. Syllable frequency and spoken word recognition: an inhibitory effect. *Psychological Reports* 119(1). 263–275. https://doi.org/10.1177/0033294116654449. http://prx.sagepub.com/content/119/1/263.abstract.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-aligner: a tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3). 192–193.

Gorman, Kyle & Daniel Ezra Johnson. 2013. Quantitative analysis. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.), *The Oxford handbook of sociolinguistics*, 214–240. Oxford, UK: Oxford University Press.

Graff, Peter. 2012. *Communicative efficiency in the lexicon*. Massachusetts Institute of Technology dissertation.

Grainger, Jonathan & Juan Segui. 1990. Neighborhood frequency effects in visual word recognition: a comparison of lexical decision and masked identification latencies. *Perception & Psychophysics* 47(2). 191–198. https://doi.org/10.3758/BF03205983. http://dx.doi.org/10.3758/BF03205983.

Hall, Kathleen Currie. 2009. *A probabilistic model of phonological relationships from contrast to allophony*. The Ohio State University dissertation.

Hall, Kathleen Currie. 2012. Phonological relationships: a probabilistic model. *McGill Working Papers in Linguistics* 22(1). 1–14.

Hall, Kathleen Currie. 2013. A typology of intermediate phonological relationships. *The Linguistic Review* 30(2). 215–275.

Hall, Kathleen Currie, Blake Allen, Michael Fry, Scott Mackie & Michael McAuliffe. 2015. *Phonological CorpusTools, Version 1.1. [Computer program]*. http://phonologicalcorpustools.github.io/CorpusTools/.

Hall, Kathleen Currie & Elizabeth Hume. Submitted. Modeling perceived similarity: the influence of phonetics, phonology and frequency on the perception of French vowels. *Laboratory Phonology*.

Hall, Kathleen Currie, Elizabeth Hume, T. Florian Jaeger & Andrew Wedel. Submitted. The message shapes phonology.

Hall, Kathleen Currie, Veronica Letawsky, Alannah Turner, Claire Allen & Kevin McMullin. 2014. Effects of predictability of distribution on within-language perception. In Santa Vīnerte (ed.), *Proceedings of the 2015 annual conference of the Canadian Linguistics As-*

sociation, 1–15. Available online at http://cla-acl.ca/wp-content/uploads/Hall_Letawsky_Turner_Alle 2015.pdf. Ottawa: Canadian Linguistics Association.

Harnsberger, James. 2000. A cross-language study of the identification of non-native nasal consonants varying in place of articulation. *The Journal of the Acoustical Society of America* 108(2). 764–783.

Harnsberger, James. 2001a. On the relationship between identification and discrimination of non-native nasal consonants. *The Journal of the Acoustical Society of America* 110(1). 489–503.

Harnsberger, James. 2001b. The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: a multidimensional scaling analysis. *Journal of Phonetics* 29(3). 303–327.

Harris, James. 1969. *Spanish phonology*. Cambridge, MA: MIT Press.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.

Heitz, Richard P. 2014. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience* 8. Article 150, 1–9.

Henrich, Joseph, Steven Heine & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33(2-3). 61–83.

Hockett, Charles. 1967. The quantification of functional load. *Word* 23(1-3). 300–320.

Holt, Lori & Andrew Lotto. 2006. Cue weighting in auditory categorization: implications for first and second language acquisition. *The Journal of the Acoustical Society of America* 119(5). 3059–3071.

Howes, Davis. 1957. On the relation between the intelligibility and frequency of occurrence of english words. *The Journal of the Acoustical Society of America* 29(2). 296–305.

Hyman, Larry. 2015. Why underlying representations? In *UC Berkeley Phonology Lab annual report*, 210–226. Available online at http://escholarship.org/uc/item/7hn3623c. Department of Linguistics, UC Berkeley.

Jaeger, T. Florian. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59(4). 434–446.

Johnson, Keith. 2005. Speaker normalization in speech perception. In David Pisoni & Robert Remez (eds.), *The handbook of speech perception*, 363–389. Malden, MA: Blackwell.

Johnson, Paul C.D. 2014. Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*. n/a–n/a. `https://doi.org/10.1111/2041-210X.12225`. `http://dx.doi.org/10.1111/2041-210X.12225`.

Jun, Jongho. 2004. Place assimilation. In Bruce Hayes, Robert Kirchner & Donca Steriade (eds.), *Phonetically based phonology*, 58–86. Cambridge, UK: Cambridge University Press.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: evidence from reduction in lexical production. *Typological studies in language* 45. 229–254.

Kataoka, Reiko & Keith Johnson. 2007. Frequency effects in cross-linguistic stop place perception: a case of /t/ - /k/ in Japanese and English. In *UC Berkeley Phonology Lab annual report*, 273–301. Available online at http://linguistics.berkeley.edu/phonlab/documents/2007/Kataoka_ Department of Linguistics, UC Berkeley.

Kaufman, Terrence. 1990. Algunos rasgos estructurales de los idiomas mayances con referencia especial al K'iche'. In Nora England & Stephen Elliott (eds.), *Lecturas sobre la*

lingüística maya, 59–114. Antigua, Guatemala: Centro de Investigaciones Regionales de Mesoamérica.

Kaufman, Terrence. 2003. *A preliminary Mayan etymological dictionary*. Ms., Foundation for the Advancement of Mesoamerican Studies. Available online at http://www.famsi.org/reports/01051/.

Keuleers, Emmanuel, Kevin Diependaele & Marc Brysbaert. 2010. Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology* 1. 174. `https://doi.org/10.3389/fpsyg.2010.00174`. `http://journal.frontiersin.org/article/10.3389/fpsyg.2010.00174`.

Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle & Marc Brysbaert. 2012. The british lexicon project: lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods* 44(1). 287–304. `https://doi.org/10.3758/s13428-011-0118-4`.

King, Robert D. 1967. Functional load and sound change. *Language*. 831–852.

Kingston, John. 1984. *The phonetics and phonology of the timing of oral and glottal events*. University of California, Berkeley dissertation.

Kingston, John. 2005a. Ears to categories: new arguments for autonomy. In Sónia Frota, Marina Cláudia Vigário & Maria João Freitas (eds.), *Prosodies: with special reference to Iberian languages*, 177–222. Berlin: Mouton de Gruyter.

Kingston, John. 2005b. The phonetics of Athabaskan tonogenesis. In *Athabaskan prosody*, 137–184. Amsterdam: John Benjamins.

Kingston, John, Joshua Levy, Amanda Rysling & Adrian Staub. 2016. Eye movement evidence for an immediate Ganong effect. *Journal of experimental psychology: Human perception and performance* 42(12). 1969–1988.

Kučera, Henry. 1963. *Entropy, redundancy and functional load in Russian and Czech*. Berlin: Mouton & Company.

Kuhl, Patricia & Paul Iverson. 1995. Linguistic experience and the "perceptual magnet effect". In Winifred Strange (ed.), *Speech perception and linguistic experience: issues in cross-language research*, 121–154. Baltimore, MD: York Press.

Kullback, S. & R.A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86.

Ladefoged, Peter & Sandra Ferrari Disner. 2012. *Vowels and consonants*. 3rd. Malden MA: Wiley-Blackwell.

Larsen, Thomas. 1988. *Manifestations of ergativity in Quiché grammar*. University of California, Berkeley dissertation.

Lindau, Mona. 1984. Phonetic differences in glottalic consonants. *Journal of Phonetics* 12. 147–155.

Lodge, Ken. 2009. *Fundamental concepts in phonology: sameness and difference*. Edinburgh: Edinburgh University Press.

Luce, Paul A. 1986. *Neighborhoods of words in the mental lexicon*. Department of Psychology, Indiana University, Bloomington, Indiana. dissertation.

Luce, Paul A. & David Pisoni. 1998. Recognizing spoken words: the neighborhood activation model. *Ear and Hearing* 19(1). 1–36.

Macario, Narciso Cojtí, Martín Chacach Cutzal & Marcos Armando Calí Semeyá. 1998. *Diccionario Kaqchikel*. Antigua, Guatemala: Proyecto Lingüístico Francisco Marroquín.

Macklin-Cordes, Jayden & Erich Round. 2015. High-definition phonotactics reflect linguistic pasts. In Johannes Wahle, Marisa Kollner, Harald Baayen, Gerhard Jager & Tineke Baayen-Oudshoorn (eds.), *Proceedings of the 6th conference on quantitative investigations in theoretical linguistics*. Tubingen, Germany. https://doi.org/10.15496/publikation-8609.

Macmillan, Neil & C. Douglas Creelman. 2005. *Detection theory: a user's guide*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates.

Maddieson, Ian. 2009. *Glottalized consonants*. Martin Haspelmath, Matthew Dryer, David Gil & Bernard Comrie (eds.). The World Atlas of Language Structures Online (WALS). Available online at http://wals.info/feature/7. Munich.

Maekawa, Kikuo. 2003. Corpus of Spontaneous Japanese: its design and evaluation. In *Spontaneous speech processing and recognition (SSPR 2003)*, paper MMO2. http://www.isca-speech.org/archive_open/sspr2003/sspr_mmo2.html. ICSA Speech Archive.

Majzul, Lolmay Filiberto Patal. 2007. *Rusoltzij ri Kaqchikel: diccionario estándar bilingüe Kaqchikel-Español*. Ciudad de Guatemala, Guatemala: Cholsamaj.

Majzul, Lolmay Filiberto Patal, Pedro Oscar García Matzar & Carmelina Espantzay Serech. 2000. *Rujunamaxik ri Kaqchikel chi': variación dialectal en Kaqchikel*. Ciudad de Guatemala, Guatemala: Cholsamaj.

de Marneffe, Marie-Catherine, John Tomlinson Jr., Marisa Tice & Meghan Sumner. 2011. The interaction of lexical frequency and phonetic variation in the perception of accented speech. In *The 33rd annual meeting of the cognitive science society [cogsci 2011]*, 3575–3580.

Martinet, André. 1952. Function, structure, and sound change. *Word* 8(1). 1–32.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, R. Harald Baayen & Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

Maxwell, Judith & Robert Hill. 2010. *Kaqchikel chronicles: the definitive edition*. Austin, TX: University of Texas Press.

McClelland, James L. & Jeffrey L. Elman. 1986. The trace model of speech perception. *Cognitive Psychology* 18(1). 1–86.

McClelland, James L., Daniel Mirman & Lori Holt. 2006. Are there interactive processes in speech perception? *Trends in cognitive sciences* 10(8). 363–369.

McClelland, James L., David E. Rumelhart & Geoffrey E. Hinton. 1986. The appeal of parallel distributed processing. In David E. Rumelhart, James L. McClelland & The PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, 3–44. Cambridge, MA: MIT Press.

McCloy, Daniel. 2014. *praat-semiauto*. https://github.com/drammock/praat-semiauto/.

McGuire, Grant. 2007. *Phonetic category learning*. The Ohio State University dissertation.

McGuire, Grant. 2010. *A brief primer on experimental designs for speech perception research*. Ms. Available online at http://people.ucsc.edu/~gmcguir1/experiment_designs.pdf.

Meyer, Julien, Laure Dentel & Fanny Meunier. 2013. Speech recognition in natural background noise. *PLoS ONE* 8(11). 1–14. https://doi.org/10.1371/journal.pone.0079279. http://dx.doi.org/10.1371%2Fjournal.pone.0079279.

Mielke, Jeff. 2012. A phonetically-based metric of sound similarity. *Lingua* 122. 145–163.

Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nelson, Noah Richard & Andrew Wedel. To appear. The phonetic specificity of competition: contrastive hyperarticulation of voice onset time in conversational English. *Journal of Phonetics*.

Newman, Rochelle S, James R Sawusch & Paul A. Luce. 1997. Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology-Human Perception and Performance* 23(3). 873–889.

Newman, Rochelle S, James R Sawusch & Paul A. Luce. 2005. Do postonset segments define a lexical neighborhood? *Memory & Cognition* 33(6). 941–960.

Norris, Dennis, James M. McQueen & Anne Cutler. 2000. Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences* 23(3). 299–325.

Oh, Yoon Mi, Christophe Coupé, Egidio Marsico & François Pellegrino. 2015. Bridging phonological system and lexicon: insights from a corpus study of functional load. *Journal of phonetics* 53. 153–176.

Oh, Yoon Mi, François Pellegrino, Christophe Coupé & Egidio Marsico. 2013. Cross-language comparison of functional load for vowels, consonants, and tones. In *Proceedings of interspeech*, 3032–3036.

Ohala, John. 1993. The phonetics of sound change. In Charles Jones (ed.), *Historical linguistics: problems and perspectives*, 237–278. London: Longman.

Peirce, Jonathan W. 2007. Psychopy–psychophysics software in python. *Journal of Neuroscience Methods* 162(1–2). 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017. http://www.sciencedirect.com/science/article/pii/S0165027006005772.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal & Emmanuel Dupoux. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* 101(3). B31–B41.

Pierrehumbert, Janet. 2001. Exemplar dynamics: word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam: John Benjamins.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Papers in laboratory phonology VII*, 101–139. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet. 2016. Phonological representation: beyond abstract versus episodic. *Annual Review of Linguistics* 2. 33–52.

Pinkerton, Sandra. 1986. Quichean (Mayan) glottalized and nonglottalized stops: a phonetic study with implications for phonological universals. In John Ohala & Jeri Jaeger (eds.), *Experimental phonology*, 125–139. Orlando: Academic Press.

Pisoni, David. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics* 13(2). 253–260.

Pisoni, David. 1975. Auditory short-term memory and vowel perception. *Memory & Cognition* 3(1). 7–18.

Pisoni, David & Jeffrey Tash. 1974. Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics* 15(2). 285–290.

Pitt, Mark A. & Arthur G. Samuel. 1993. An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance* 19(4). 699–725.

Port, Robert & Adam Leary. 2005. Against formal phonology. *Language* 81(4). 927–964.

Quené, Hugo & L.E. van Delft. 2010. Non-native durational patterns decrease speech intelligibility. *Speech Communication* 52(11fffdfffdfffd12). Non-native Speech Perception in Adverse Conditions, 911–918. https://doi.org/http://dx.doi.org/10.1016/j.specom.2010.03.005. http://www.sciencedirect.com/science/article/pii/S0167639310000580.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Version 3.0.1, retrieved from http://www.R-project.org/. R Foundation for Statistical Computing. Vienna, Austria.

Redford, Melissa A. & Randy L. Diehl. 1999. The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *The Journal of the Acoustical Society of America* 106. 1555–1565. https://doi.org/10.1121/1.427152.

Renwick, Margaret. 2014. *The phonetics and phonology of contrast: the case of the Romanian vowel system*. Berlin: Walter de Gruyter.

Repp, Bruno H & Robert G Crowder. 1990. Stimulus order effects in vowel discrimination. *The Journal of the Acoustical Society of America* 88(5). 2080–2090.

Rice, Glenn A. & David Owen Robinson. 1975. The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition* 3(5). 513–518. https://doi.org/10.3758/BF03197523. http://dx.doi.org/10.3758/BF03197523.

Richards, Michael. 2003. *Atlas lingüístico de Guatemala*. Instituto de Lingüístico y Educación de la Universidad Rafael Landívar.

Rose, Sharon & Lisa King. 2007. Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* 50(4). 451–504.

Russell, Susan. 1997. *Some acoustic characteristics of word initial pulmonic and glottalic stops in Mam*. Simon Fraser University MA thesis.

Sakoe, Hiroaki & Seibi Chiba. 1971. A dynamic programming approach to continuous speech recognition. In *Proceedings of the seventh international congress on acoustics*, vol. 3, 65–69.

Sebastián-Gallés, Núria. 2005. Cross-language speech perception. In David Pisoni & Robert Remez (eds.), *The handbook of speech perception*, 546–566. Malden, MA: Blackwell.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Silverman, Daniel. 2006. *A critical introduction to phonology: of sound, mind, and body*. London & New York: Continuum.

Silverman, Daniel. 2012. *Neutralization*. Cambridge, UK: Cambridge University Press.

Smits, Roel, Joan Sereno & Allard Jongman. 2006. Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance* 32(3). 733–754.

Steriade, Donca. 2001. Directional asymmetries in place assimilation: a perceptual account. In Keith Johnson & Elizabeth Hume (eds.), *The role of speech perception in phonology*, 219–250. New York: Academic Press.

Steriade, Donca. 2009. The phonology of perceptibility effects: the p-map and its consequences for constraint organization. In Kristin Hanson & Sharon Inkelas (eds.), *The nature of the word: studies in honor of Paul Kiparsky*, 151–179. Cambridge, MA: MIT Press.

Stevenson, Sophia & Tania Zamuner. 2017. Gradient phonological relationships: evidence from vowels in French. *Glossa* 2(1). 1–22.

Surendran, Dinoj & Partha Niyogi. 2003. *Measuring the usefulness (functional load) of phonological contrasts*. Tech. rep. Technical Report TR-2003. Chicago: Department of Computer Science, University of Chicago.

Surendran, Dinoj & Partha Niyogi. 2006. Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In Ole Nedergaard Thomsen (ed.), *Competing models of linguistic change: evolution and beyond*, 43–58. Amsterdam: John Benjamins.

Sze, WeiPing, SusanJ. Rickard Liow & MelvinJ. Yap. 2014. The chinese lexicon project: a repository of lexical decision behavioral responses for 2,500 chinese characters. English. *Behavior Research Methods* 46(1). 263–273. `https://doi.org/10.3758/s13428-013-0355-9`. `http://dx.doi.org/10.3758/s13428-013-0355-9`.

Tang, Kevin. 2015. *Naturalistic speech misperception*. University College London dissertation.

Tang, Kevin, Ryan Bennett & Juan Ajsivinac Sian. In preparation. Contextual predictability influences word duration in a morphologically complex language (Kaqchikel Mayan).

Tang, Kevin, Ryan Bennett & Juan Ajsivinac Sian. 2015. *Modelling Phonetic and Phonological Variation with 'Small' Data: Evidence from Kaqchikel Mayan*. Presentation at *Laboratory Phonology 15* conference.

Tang, Kevin & Andrew Nevins. In prep. A graceful degradation account of lexical retrieval: evidence from naturalistic misperception.

Tang, Kevin & Andrew Nevins. 2014. Measuring segmental and lexical trends in a corpus of naturalistic speech. In Hsin-Lun Huang, Ethan Poole & Amanda Rysling (eds.), vol. 2, 153–166. Amherst, MA: GLSA.

Tilsen, Sam. 2016. Selection and coordination: the articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55. 53–77.

Trubetzkoy, Nikolai. 1939. *Grundzüge der phonologie*. English translation published 1969 as *Principles of phonology*, trans. C.A.M. Baltaxe. Berkeley: University of California Press. Travaux du cercle linguistique de Prague.

van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* 67(6). 1176–1190. `https://doi.org/10.1080/17470218.2013.850521`.

Vitevitch, Michael. 2002. Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and speech* 45(4). 407–434.

Vitevitch, Michael & Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40(3). 374–408.

Vitevitch, Michael & Paul A. Luce. 2016. Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics* 2. 75–94.

Wang, Marilyn D. & Robert C. Bilger. 1973. Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America* 54(5). 1248–1266. https://doi.org/10.1121/1.1914417.

Wedel, Andrew. 2004. *Self-organization and categorical behavior in phonology*. UC Santa Cruz dissertation.

Wedel, Andrew, Scott Jackson & Abby Kaplan. 2013a. Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and speech* 56(3). 395fffdfffdfffd417.

Wedel, Andrew, Abby Kaplan & Scott Jackson. 2013b. High functional load inhibits phonological contrast loss: a corpus study. *Cognition* 128(2). 179–186.

Werker, Janet & John Logan. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics* 37(1). 35–44.

Werker, Janet & Richard Tees. 1984a. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant behavior and development* 7(1). 49–63.

Werker, Janet & Richard Tees. 1984b. Phonemic and phonetic factors in adult cross-language speech perception. *The Journal of the Acoustical Society of America* 75(6). 1866–1878.

Whalen, D.H. & Joyce McDonough. 2015. Taking the laboratory into the field. *Annual Review of Linguistics* 1(1). 395–415. https://doi.org/10.1146/annurev-linguist-030514-124915. http://dx.doi.org/10.1146/annurev-linguist-030514-124915.

Wright, Charles E. 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition* 7(6). 411–419. https://doi.org/10.3758/BF03198257.

Wright, Richard. 2004. A review of perceptual cues and cue robustness. In Bruce Hayes, Robert Kirchner & Donca Steriade (eds.), *Phonetically based phonology*, 34–57. Cambridge, UK: Cambridge University Press.

Wright, Richard, Sharon Hargus & Katharine Davis. 2002. On the categorization of ejectives: data from Witsuwit'en. *Journal of the International Phonetic Association* 32(1). 43–77.

Xu, Yi. 2010. In defense of lab speech. *Journal of Phonetics* 38(3). 329–336.

Yao, Yao. 2011. *The effects of phonological neighborhoods on pronunciation variation in conversational speech*. University of California, Berkeley dissertation.

Yap, Melvin J, Daragh E Sibley, David A Balota, Roger Ratcliff & Jay Rueckl. 2015. Responding to nonwords in the lexical decision task: insights from the english lexicon project. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(3). 597.

Yarkoni, Tal, David Balota & Melvin Yap. 2008. Moving beyond Coltheartfffdfffdfffds N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review* 15(5). 971–979.

Yu, Alan C.L. 2011. On measuring phonetic precursor robustness: a response to Moreton. *Phonology* 28(3). 491–518.

Zipf, George Kingsley. 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.