To appear in *Cognition*

Tone slips in Cantonese: Evidence for early phonological encoding

John Alderete, Queenie Chan, Henny Yeung

Simon Fraser University

Abstract. This article examines speech errors in Cantonese with the aim of fleshing out a larger speech production architecture for encoding phonological tone. A corpus was created by extracting 2,462 speech errors, including 668 tone errors, from audio recordings of natural conversations. The structure of these errors was then investigated in order to distinguish two contemporary approaches to tone in speech production. In the tonal frames account, tone is encoded like metrical stress, represented in abstract structural frames for a word. Because tone cannot be mis-selected in tonal frames, tone errors are expected to be rare and non-contextual, as observed with stress. An alternative is that tone is actively selected in phonological encoding like phonological segments. This approach predicts that tone errors will be relatively common and exhibit the contextual patterns observed with segments, like perseveration and anticipation. In our corpus, tone errors are the second most common type of error, and the majority of errors exhibit contextual patterns that parallel segmental errors. Building on prior research, a two-stage model of phonological tone encoding is proposed, following the patterns seen in tone errors: Tone is phonologically selected concurrently with segments, but then sequentially assigned after segments to a syllable.

Keywords: speech errors, phonological encoding, speech production, tone, similarity, activation dynamics

1. Introduction

The majority of the world's languages are tone languages (Yip, 2002). In these languages, tone structures exhibit distinctive pitch shapes that can make contrasts in otherwise identical words. Producing a word accurately in a tone language involves correctly selecting tone from a larger tonal inventory. For example, producing the word 'father' in Mandarin involves pairing up the segmental string *ba* with the falling tone [51]: *ba51*, and the mis-selection of tone produces a completely different word, e.g., *ba55* 'eight'.[1] Despite the prevalence of tone, several empirical and theoretical questions remain unanswered about the nature of tone encoding in speech production.

Many of the conclusions about the encoding of tone have been made on the basis of speech errors. Early work on tone production errors focused on the psychological reality of tone and its status as a speech-planning unit. Investigation of speech errors in Mandarin (Moser, 1991; Shen, 1993) and Thai (Gandour, 1977) supported the view that tone is a viable planning unit that

[1] Here and throughout we use the Chao system of transcribing tone which approximates a pitch level on a numerical scale from 1 (low) to 5 (high) and pitch contours through tone onsets and offsets, e.g., [51] is a fall in pitch from the highest point to the lowest (Chao, 1930).

can be perseverated, anticipated, and exchanged in form encoding processes, much like speech error patterns with consonants and vowels found in Indo-European languages like English and German (Berg, 1988; Fromkin, 1971). For example, in the Mandarin error, ***chang55*** *an55 jie55* 'Avenue of Heavenly Peace' (Intended: ***chang35*** *an55 jie55*, from Moser (1991:7)), the first -syllable has a level high tone [55] rather than the correct rising tone [35], which is likely an error that anticipates the tone of the next syllable.

Subsequent work on Chinese languages expanded questions about the nature of the representation and encoding of tone in speech production with larger datasets. From this literature, two broader issues remain unresolved. The first concerns whether tone is actively selected in phonological encoding, or an inherent property of the form representation that is not actively selected. Contemporary accounts of tone encoding, and later work developed from them, are divided on precisely this issue. In one set of analyses, tone is selected in phonological encoding using a mechanism similar to the one for selecting phonological segments (Wan & Jaeger, 1998 et seq.). An alternative view assumes that information about tone is not actively selected but rather mapped from a lemma representation to a prosodic frame, as many assume metrical stress is encoded (J.-Y. Chen, 1999; Roelofs, 2015).

The second issue concerns the role of 'proximate units' of speech encoding, a debate that asks whether word-form encoding in Chinese lends greater significance to the syllable in speech planning, as opposed to the segment in Indo-European languages. While our analysis of tonal speech errors does not address the central question of this debate about proximate units—the primary focus is the first issue concerning tone encoding—our work does bear on the question of whether tones, segments, or syllables slip more often when making speech errors. Below, we provide further background to each of these questions in turn.

1.1 Early or late encoding of tone and interactivity

In one of the first systematic studies of speech errors in a Chinese language, Wan and Jaeger (1998) examined 788 speech errors in Mandarin, including 83 tone errors. Their results suggested that tone involves a selection mechanism in phonological encoding similar to segments, but segments and tone are not selected together as integral wholes. Specifically, Wan and Jaeger argued that tone encoding is distinct from segment encoding because both segment and tone errors occur with some regularity, but they tend not to occur in the same utterances. Further, certain errors (sequential blends) show that tones may be retained when the rhymes associated with them are deleted, e.g., Intended: /tɕjɛn21 fɑŋ55 pan51 an51/ → tɕjɛn21 **pan55 …** 'prosecutor dealt with this case' (p. 442). At the same time, tone was also assumed by Wan and Jaeger to be encoded at the same operational level as segments, because tone participates in the same types of contextual errors that involve mis-ordering of units (e.g., perseverations, anticipations, and exchanges). In addition, Wan and Jaeger found that tone structure appears to be part of the phonological structure of the lexicon, which again supports the hypothesis of active selection of tone in phonological encoding. Lexical substitutions, for example, had a greater-than-chance probability of bearing the same tone in the intended and error words. Adopting a production model similar to Bock and Levelt (1994), based on earlier and related ideas from Garrett (1984), Wan and Jaeger place encoding of both segments and tone in word-form retrieval where phonological units are retrieved for already activated lemmas. We dub this approach as "early encoding" because tone is the result of speech planning in phonological encoding, rather than a late implementation process.

Another large-scale study of Mandarin came to rather different conclusions about how tone is encoded. J.-Y. Chen (1999) examined 987 speech errors, but found only 24 tone errors, and argued further that many of these cases could be attributed to other, non-tone error processes. Chen argued that the relative rarity of tone errors is uncharacteristic of segmental errors, which tend to be the most common type in speech error corpora in well-known Indo-European languages (Shattuck-Hufnagel, 1979; Stemberger, 1982/1985), as well as in Chinese languages like Mandarin (J.-Y. Chen, 1999; Wan & Jaeger, 1998). Instead, Chen proposed that the role of tone is more akin to the role of metrical stress in speech production. Stress errors are also exceedingly rare, with some cases of putative stress errors re-analyzed as the mis-selection of morphologically related words (Cutler, 1980). Based on the analogy with stress, Chen argues that tone is not actively selected in word-form retrieval like segments. Instead, tone is an inherent property of a word form that is mapped from lemmas to structural frames of words, as is the case for metrical structure in WEAVER++ (Levelt, Roelofs, & Meyer, 1999). A recent adaptation of WEAVER++ to Mandarin Chinese illustrates how tone is mapped to a tonal frame (Roelofs, 2015). In this account (see Figure 1), as with the WEAVER++ account of metrical stress, tone is represented diacritically with special labels in structural frames in phonological encoding and implemented as an explicit tone structure at the later level of phonetic spell-out. This approach contrasts with the early active selection of tone in phonological encoding on Wan and Jaeger's account. On the late encoding approach, the reason tone errors are rare is that tone is not actively selected from a larger tonal inventory in phonological encoding and so, like stress, it cannot be mis-selected.

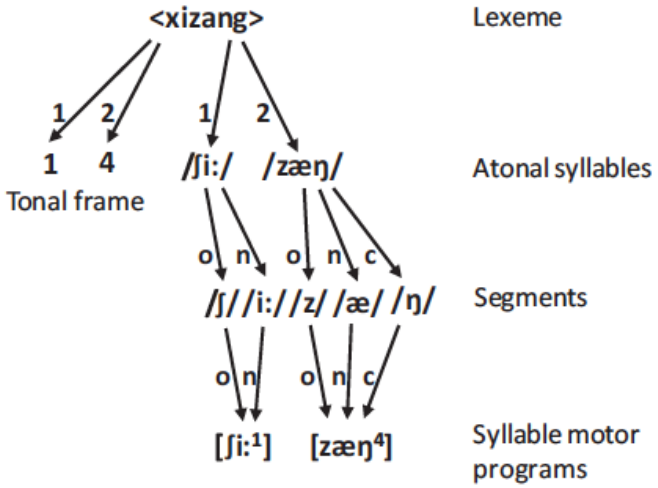

Figure 1. Roelofs' (2015) representation of the the WEAVER++ model for speech production in Mandarin, where tone is represented as part of the lexicon, but does not combine with the segmental components of the syllable until the level of 'Syllable motor programs,' where the superscripts indicate when tone is specified in production planning.

To see how these two approaches have contributed to current research on tone, it is important to underscore the fundamental differences between them. In the early encoding approach advocated in Wan and Jaeger (1998), tone errors pattern empirically with segmental errors (relatively common), and thus tone is actively selected in word form retrieval. In the late encoding approach, tone errors are unlike segmental errors but like stress, and therefore are treated like metrical stress in word form retrieval: Tone is inherent to the structural frames of words but not actively selected. These two fundamentally different accounts have supported

active debate in current work on the encoding of tone.

The latter view, that tone is inherent to structural frames, is wide-spread and taken as a given in both theoretical models and empirical investigations. For example, a number of current models of word-form encoding developed for Chinese languages follow Chen (1999) in that they posit abstract structural frames with tonal information (J.-Y. Chen, Chen, & Dell, 2002; J.-Y. Chen & Dell, 2006; Roelofs, 2015). The assumed motivation for this position is again that tone cannot be mis-ordered in structural frames, and so tonal frames account for the apparent rarity of tone errors (see e.g., Chen, Chen, & Dell (2002: 769)). This assumption contrasts with encoding segments, which are actively selected and therefore can be mis-ordered. O'Seaghdha, Chen, and Chen (2010) likewise posit word-shape frames to which tone is assigned in the context of the proximate unit hypothesis (see Figure 2 and ensuing discussion). Though the use of structural frames in this model suggests a treatment of tone like metrical stress, it should be pointed out that the specific mechanism for assigning tone is not the focus of this theory, and it can be extended to include tone selection, as we show in Section 4.3 below. In a further development, O'Seaghdha (2015) points out that the existence of speech errors involving substitutions of entire syllables without tone (documented in J.-Y. Chen (2000)) shows that syllables are selected prior to tone association, further strengthening the parallel with stress, because segments in languages like English are selected prior to association to metrical positions.

In addition to these theoretical contributions, the account given in Chen (1999) that tone slips are extremely rare has had tremendous impact in guiding empirical investigations of tone processing. For instance, motivated by the apparent under-representation of tone errors, Kember, Croot, and Patrick (2015) designed a tone twister experiment that was structured to distinguish segmental and tone errors in terms of their overall frequencies and the contexts in which they arise. In an fMRI study of the neural correlates of segmental and tone in Mandarin, Gandour et al. (2003) used the assumed disparity between segmental and tonal errors to motivate a hypothesis in which tone is associated with syllable internal units rather than whole syllables. In a different domain, Simner, Hung, and Shillcock (2011) use the corollary between lexical stress in English and tone in Mandarin to test whether tone, like stress, is associated with color in Chinese-speaking grapheme-color synaesthetes. The apparent rarity of tone slips has also been used to interpret experimental findings. For example, Chang, Lee, Tzeng, and Kuo (2014) use the rarity of tone errors to reject an alternative explanation for longer reaction times with sequences of Mandarin tone 3 based on their inherent difficulty. In sum, Chen's (1999) account of tone slips has had considerable impact on the empirical generalizations that drive much work on tone.

Against this background, a cross-current of research has developed that supports the contention that tone is encoded like segments, as proposed in Wan & Jaeger (1998) and earlier work (Fromkin, 1980; Moser, 1991; Shen, 1993). Wan (2006) examined 876 tone errors in patients with aphasia and concluded that the impairment of tone is comparable to the impairment of consonants, thus showing a similar underlying architecture. Likewise, Liu and Wang (2008) investigated tone and segmental errors in Taiwanese (a southern Min language related to Mandarin, but with a distinct tonal inventory) and found that both tone and segmental errors are contextually conditioned, and therefore incompatible with the idea that the encoding of tone is inherent to structural frames (although tone and segment errors did differ with respect to other factors, like the directionality of source elements). A variety of neurological studies have also investigated tone processing that lend support to the early encoding approach. This research contrasted right-lateralization associated with pure (non-linguistic) tones, music, and prosodic

intonation with left-lateralization of language processing in linguistic contexts, and found strong evidence of left-lateralized lexical tone but no evidence for right lateralization (Gandour, 1998; Packard, 1986; Van Linker & Fromkin, 1973). In an event-related potential study, Brown-Schmidt and Conseco-Gonzalez (2004) found that anomalous sentences deriving from tone structure elicited a robust N400 effect and concluded that, like this prior work, lexical tones are processed as linguistic information and not as pure tones or intonational prosody. These authors engage directly with the speech error data reported in Chen (1999), and while they note these findings are not incompatible with processing tone like prosody, they nonetheless identify important differences between lexical tone in languages like Mandarin and lexical stress in English that are consistent with the contention that tone is processed as linguistic information rather than non-linguistic information or intonation.

Another important observation supportive of early encoding is Wan and Jaeger's (1998) finding that error words in Mandarin lexical substitutions tend to share a tone with the intended word. In the present paper, we examine feedback effects such as these, because their existence is central to theoretical claims of whether tone interacts with other phonological units in speech planning. Both backward and forward feedback effects have been documented in a number of different ways, but all generally involve increases in error rates when the intended and pronounced words share certain form elements. We refer to these phenomena collectively as interactive spreading effects, following Dell (1986, 1988), because they depend on spreading activation across different levels of linguistic structure.

One well-known example of interactive spreading is the repeated phoneme effect, in which a shared sound in two neighboring words increases the rate of errors involving other sounds contained in those words (Dell, 1984; MacKay, 1970). For example, the two words of the phrase *deal beak* share the vowel phoneme [i], and this shared structure is hypothesized to influence the production of *deal* such that it increases the chance of a *d* to *b* substitution. Interactive spreading effects are thought to stem from the activation dynamics of selecting words and sounds. In a variety of language production models employing such a dynamics (Dell, 1986; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Goldrick & Blumstein, 2006; Stemberger, 1982/1985), words and sounds are selected over others because they have a higher value of activation. Because *deal* and *beak* share [i], anticipatory activation of *beak* increases the flow of activation between the two words, which in turn increases the possibility of *b* intruding on *d* in the first word (Dell, 1988). While this interactivity is not espoused by all models (e.g., Levelt et al. (1999)), interactive spreading effects provide clear evidence for early integration because they require explicit representation of form elements in both phonological and grammatical encoding, and spreading activation across these linguistic levels. We provide three examples of interactive spreading effects involving tone in Section 3.3.

Our focus here is on documenting spreading interactive effects in support of model development, but not on adjudicating between correct and incorrect models of language production. In particular, we extend an existing model of phonological encoding based on the proximate unit hypothesis (O'Seaghdha et al., 2010) in Section 4.3 because we believe it can be naturally extended to account for the observed facts. This extension is not intended to rule out other models. Thus, discrete feedforward models have been argued to account for some of the error biases, like the lexical bias, which is commonly argued to support interactive spreading with feedback from speech comprehension (Roelofs (2004), cf. Rapp & Goldrick (2000)), and which leaves open the possibility that interactive spreading effects in encoding tone could be accounted for in these models with related mechanisms. However, the assumption that tone is

encoded like metrical stress (see Tonal Frame in Figure 1) may be problematic. This assumption is made specifically to account for the rarity of tone errors because tones specified on a prosodic frame do not interact with other tones in their local context (J.-Y. Chen, 1999). The present study will provide a wealth of facts about the interaction of encoding tone with other elements.

To sum up, the distinction between these accounts of tone in Chinese speech production boils down to either the active selection of tone in an early interactive encoding process (Wan & Jaeger, 1998), or a non-interactive account similar to stress, implemented as a mapping of tone to a prosodic frame and a later phonetic spell-out (J.-Y. Chen, 1999). A critical way of adjudicating between these two approaches is to ask whether tone is explicitly represented and selected in word-form encoding, and if so, we expect to find a non-trivial number of tone mis-selections due to the surrounding context, as found with segmental errors. If, on the other hand, tone is not actively selected, then we would not expect tone mis-selections that are contextually linked. Like stress errors, the small number of apparent tone mis-selections might then be classified as other kinds of errors, for example, blends or lexical substitutions involving different tones (J.-Y. Chen, 1999). Our investigation below provides clear evidence for the former view, which we revisit in Section 4 (Discussion).

1.2 Tone interactions with 'proximate units' in Chinese speech planning

A second major debate in the Chinese speech production literature centers around the 'proximate unit' in speech planning. A prominent view is that atonal syllables—and not segments or tones—are central units of speech production planning (J.-Y. Chen et al., 2002; T.-M. Chen & Chen, 2013; O'Seaghdha et al., 2010; Roelofs, 2015). As illustrated in Figure 2, this account suggests that Chinese lexemes are first accessed as syllable-sized chunks without tonal specification in the speech planning process. Moreover, lexemes are not accessed segment-by-segment either, as is argued to be the case for Indo-European languages, like English. This hypothesis does not preclude downstream segmental effects, as syllables can be decomposed further into segmental units in later stages of speech production and phonetic spell-out (also shown in Figure 2), but it does suggest that syllables are the most "proximate unit" for speech planning in the sense that it is accessed immediately after lemma selection.

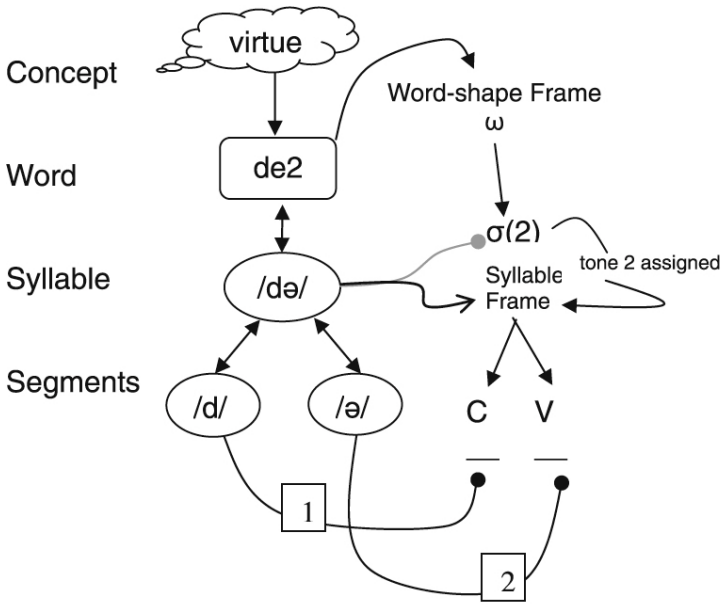

Figure 2. Form encoding with syllable as proximate unit (from O'Seaghdha et al. 2010)

The bulk of evidence for the proximate unit hypothesis comes from implicit priming, or form preparation, studies which evaluate naming latencies for lists of multiple words that have varying degrees of overlap with respect to different speech units (J.-Y. Chen et al., 2002; O'Seaghdha et al., 2010). Here, native Mandarin speakers show faster naming latencies for multiple words that shared common syllables (without necessarily the same tone), but latencies for words that overlapped in sub-syllabic information, like having common onsets, were not faster. These effects are striking, not only in that they are so different than parallel results for speakers of Indo-European languages, where this implicit priming of just the initial consonant can rapidly speed naming (Meyer, 1990), but also that they show no priming effects unless the whole syllable in Mandarin (irrespective of tone) is intact.

Converging evidence for the proximate unit hypothesis comes also from A. W.-K. Wong and Chen (2008), who used a picture-word interference paradigm with Cantonese stimuli in which naming latencies to a visual stimulus were not affected by an auditory distractor, unless that distractor also had the same onset and rhyme (i.e., was an intact syllable without necessarily the same tone). This effect on latencies was not found when the distractor shared just an onset, a rhyme, or a tone (or even when sharing the onset *and* tone). However, A. W.-K. Wong and Chen (2008), as well as several follow-up studies that Wong and colleagues conducted (A. W.-K. Wong & Chen, 2009, 2015; A. W.-K. Wong, Huang, & Chen, 2012), suggested that there are circumstances under which sub-syllabic overlap is sufficient to induce an interference effect on naming latencies, an apparent contradiction to the hypothesis that the atonal syllable is the proximate unit of Chinese speech production. Specifically, partial overlap of the vowel and coda, the onset and vowel, the onset and coda, as well as overlap of tone information in some combinations with these segments, have a similar effect as the intact syllable.

More recent results from other experimental paradigms have offered general support of the proximate unit hypothesis. For example, a masked priming paradigm, where the prime was not auditory, but rather a masked visual stimulus, showed effects only from an intact syllable, and not from segments (J.-Y. Chen, O'Seaghdha, & Chen, 2016). One event-related potential (ERP) study replicated null effects of segmental priming in behavior, but did show segmental

priming in ERPs (Qu, Damian, & Kazanina, 2012), while other ERP studies have reinforced the idea that syllables are a salient planning unit (Wang, Wong, Wang, and Chen (2017); A. W.-K. Wong, Chiu, Wang, Wong, and Chen (2018). Recent work suggests ERP indices for both syllable and segment retrieval are observable, with syllable-related components occurring prior to phoneme-related components in time (Qu et al., 2012; Zhang & Damian, 2019).

All of this work has relied upon experimental tasks to study the nature of proximate units in Chinese speech planning, and so here we examined how the study of naturalistic speech errors may shed light on this issue. For example, a strong version of the proximate unit hypothesis might assume that single segmental or single tone errors should be extremely rare, and could rather be explained by the retrieval of an incorrect syllable. The syllable structure of Chinese languages is relatively restrictive, and while it is difficult to identify unambiguous segment or syllable errors (i.e., single segmental errors often result in legal syllables in Chinese), this question has not yet been extensively explored in the speech error literature (cf., J.-Y. Chen (2000)). This issue thus comprises a second region of inquiry for our data, which offers a new way of probing these questions using naturalistic speech productions (rather than laboratory-induced priming effects). Here we ask what speech errors in Chinese might say about the proximate unit hypothesis, exploring whether unambiguous segmental and tone errors can be distinguished from errors that could occur just at the syllabic level.

1.3 Summary of the current study

This report is a large-scale study of speech errors in Cantonese extracted from natural conversations. It documents 2,462 speech errors, including 668 errors involving tone. Here we assess the pattern of tone errors in Cantonese to determine whether their relative frequency and distribution support an account of early active selection of tone in word-form retrieval or inherent tone in structural frames followed by later implementation in articulation. Specifically, we examine several facets of tone errors that relate to this encoding question, including tone in complex speech errors involving other structures, if tone is retained in sequential blends, if tone encoding interacts with the encoding of other structures, and whether there are similarity effects in single tone confusions. We also consider the role of tones, segments, and syllables in production, asking whether our database can yield insight about the relative frequency of errors for each of those units, and whether there are unambiguous errors associated with each unit type.

In this investigation, we aim to contribute both to the empirical understanding of tone processing and to language production in Asian languages in general, which have been historically understudied (Costa, Alario, & Sebastián-Gallés, 2007; Griffin & Crew, 2012). This work is the first comprehensive study of speech errors in Cantonese, and the largest speech error collection to date of any Asian language. We conclude by proposing more explicit mechanisms for the encoding of tone in Chinese languages. This discussion also contributes to broader issues in language production concerning the nature of interactivity of encoding processes, the proper treatment of prosodic structure in language processing, and the serial order problem in phonological encoding.

2. Methods

2.1 Definition of speech error

We followed standard practice in the field by defining a speech error as “an unintended, nonhabitual deviation from a speech plan” (Dell, 1986: 284). This definition encompasses sound

errors and word errors of various types (substitutions, deletions, additions, shifts), different sources from context (perseverations, anticipations, etc.), and also some morpho-syntactic and syntactic errors, like sentence blends and functional role mis-selections. Under this definition, speech errors are not false starts, idiolectal or dialectal variants, changes to a speech plan, or patterned variation, like casual speech or habitual linguistic patterns (Bauer & Benedict, 1997; Cheung, 1986; Matthews & Yip, 2011). The corpus that we used, the SFU Speech Error Database (SFUSED) Cantonese (Alderete & Chan, 2018), also includes phonetic errors (correctly selected sounds that are mis-articulated), which are distinct from phonological errors (mis-selected sounds that are correctly articulated), because of the increasing importance of these error types in speech analysis (Frisch & Wright, 2002; Goldrick & Blumstein, 2006). However, we only analyzed phonological sound errors in this study, following prior work on tone errors.

2.2 The corpus: SFUSED Cantonese 1.0

Speech errors were collected from audio recordings by a team of four trained data collectors. In particular, data collectors listened to 1,917 minutes (roughly 32 hours) of natural conversations from 50 podcast episodes. The podcasts came from three different podcast series in which commentators and guests discussed entertainment topics from contemporary film and television, and also lifestyle topics concerning interpersonal relationships. These podcast series were chosen because they had high production quality, a balance of speakers for age and gender, and long intervals of unscripted speech. These recordings did contain some scripted material, like set introductions and commercials, but speech errors were not collected from these portions of the recording, and they are not factored in the total minutes given above.

The four data collectors were native speakers of Cantonese, and also fluent (n = 3) or semi-fluent (n = 1) in English. Three of them were advanced undergraduate students at Simon Fraser University, and already had a strong background in linguistics. The fourth data collector (second author) was a graduate student in linguistics during data collection and later became a data analyst, overseeing the management and classification of the data. In addition to their general background in linguistics, data collectors were trained to detect and analyze speech errors. This training began with a one-hour introduction to speech errors that included a variety of examples from English illustrating valid and invalid errors. After the introduction, trainees were asked to spend an hour listening for errors in their daily lives to illustrate the viability of errors in natural conversations. Following a discussion of the errors collected (but discarded) from this exercise, data collectors were then given a set of three listening tests designed to enhance their capacity for detecting errors in English before moving on to Cantonese. In each listening test, trainees were asked to detect all the errors in a 30-minute recording that had been pre-screened for errors. Data collectors submitted their observed errors to the first author, who gave feedback on both valid and invalid errors, and also the errors in the recording that they missed. Trainees were also given two phonetic exercises designed to assess accuracy in transcribing Cantonese speech in phonetic notation and tone classification.

After training, the data collectors applied the concepts learned in their training to Cantonese speech. Each of the 50 podcast episodes were examined independently by two data collectors, who recorded their observed errors in spreadsheets and submitted them to the database manager (second author). Each error was then re-examined by an analyst to verify that the error met the definition of an error given above, e.g., weeding out changes of the speech plan, habitual variants of a form, etc. In particular, 3,877 errors were submitted by data collectors

(from 1,917 minutes of natural speech), and of these, 1,353 (roughly 35%) were excluded because they did not meet the definition of an error.

The collection of speech errors can be plagued by data reliability problems and problems related to the fact that human data collectors are affected by perceptual biases when collecting errors (Bock, 1996; Pérez, Santiago, Palma, & O'Seaghdha, 2007). We feel that the composition of our corpus reflects good methodological decisions that mitigate these factors to produce a reliable and robust dataset. In particular, after excluding the "false positive" errors discussed above, 2,462 valid errors were detected from 1,917 minutes of speech, which means that a valid error was detected, on average, once every 46.7 seconds. The corpus thus reflects a better sample of the true population of speech errors than other studies that do not use audio recordings and do not use pairs of trained listeners. As documented in Alderete and Davies (2018), prior studies have detection rates (an error on average every 5 to 6 minutes) that undershoot our detection rates by a wide margin. Our corpus also has a relatively low rate of uncommon but highly salient errors, like sound exchanges, and higher rates of phonotactic violations, which are further indicators of higher data reliability (Alderete & Tupper, 2018).

2.3 Classification of errors

The speech errors in our corpus have been categorized within a standard taxonomy that cross-classifies errors by linguistic unit, type of operation, and direction in contextual errors (Dell, 1986; Shattuck-Hufnagel, 1979; Stemberger, 1993), a taxonomy that is commonly applied to Chinese languages (J.-Y. Chen, 1999; Shen, 1993; Wan & Jaeger, 1998). A classification within this taxonomy involves establishing an intended sound or word, an intruder that supplants it, and possibly a source unit identical to the intruder in a neighboring word. For example, in the phonological substitution error in Table 1, the intended sound [g] of the intended word [**g**am25a:22] is supplanted by the intruder [dz], which occurs downstream in two source words (throughout error words are prefixed with a "/" and source words with a "^"). This is a contextual error because its context includes words that contain a source sound or word. The lexical substitution also shown in Table 1 lacks a source for the intruder word, so it is non-contextual.[2]

[2] Here and throughout we use an adapted version of the International Phonetic Alphabet to transcribe consonants and vowels, except we follow Jyutping and Yale romantization convention and transcribe the contrast between aspirated versus unaspirated sounds as voiceless versus voiced sounds; for example, the difference between [b/p] in our system is really [p/p^h] in the IPA. Tone is transcribed using the Chao tone transcription system, as explained in footnote 1 and in Section 2.4.

Table 1. Contextual and non-contextual errors

Phonological substitution, contextual: anticipation
咁好喇, /怎就# 到^最後 ^就吳君如就想辦法... gam25 hou25 la:33, /**dz**am25a:22 dou33 ^dzœi33hau22 ^dzau22 m21gwan55jy21 dzau22 sœŋ25 ba:n22fa:t33 … (Intended: gam25a:22 'so')
'That's good, so in the end Sandra Ng thought of a solution …'
Lexical substitution, non-contextual
應該去到嗰個位,/觀眾# 都係,吓,原來係咁樣㗎. jiŋ55goi55 hœi33dou33 go25 go33 wai25, /**gun55dzuŋ33** dou55 hai22, ha:25, jyn21loi21 hai22 gam25jœŋ25 ga:21 (Intended: tiŋ33dzuŋ33 'listeners')
'When you get to that part, the listeners should be like, oh, so that's how it is.'

With this taxonomy, speech errors can be broken down by type (i.e., substitutions, additions, deletions, and shifts) as well as by the linguistic unit affected by these operations. The units relevant for our discussion are segments (or strings of segments), tones, morphemes, words, phrases, and importantly, syllables, given their importance in production planning (J.-Y. Chen, 2000; J.-Y. Chen et al., 2002). Errors can be further cross-classified by direction: perseverations (source precedes the error), anticipations (source follows the error), exchanges (intended word and a source word exchange places), and combined perseverations and anticipations. Some errors are ambiguous in the sense that they could be classified as more than one type (e.g., a sound error that results in a lexical word could be a lexical substitution). In such contexts, we follow Stemberger (1982/1985) in employing Occam's Razor to argue for the most likely classification, while retaining alternative classifications as part of the record in case the ambiguous status needs to be re-examined. All of the various options for classification are illustrated in Table 3 below, as well as more complex errors that combine more than one error type within a single example.

2.4 Cantonese tone

All canonical syllables in Cantonese words have one of the six tones shown below. We follow Chen's (2000) analysis of tone targets, which can be cross-classified by register (high or low pitch range) and type (pitch shape). The numerals suffixed to each syllable in the examples use the Chao transcription system which approximates the pitch shapes of these tones (Chao, 1930, 1947), adapted here for Modern Cantonese. We also show the corresponding Jyutping tone number, another common way of marking tone used by the Linguistic Society of Hong Kong. The Chao transcription and tonal types are relevant to the analysis of similarity, which we examine below in the discussion of tone substitutions.

Table 2. Cantonese tonal inventory

Register	Type	Tone targets	Jyutping	Example (with tone suffix)
high	level	H	1	wan55 'warm' (cf. wan25), also: wat5 'twisted'
high	level	M	3	wan33 'to shut/lock up', also: wa:t3 'to dig'
high	rising	MH	2	wan25 'to look for'
low	level	L	6	wan22 'to transport', also: wat2 'pit (of fruit)'
low	rising	LM	5	wan23 'to allow'
low	falling	ML	4	wan21 'cloud'

In some varieties of Cantonese, the high level tone [55] is realized in some words as a high falling tone [53] or in free variation with this tone, but in the standard variety, and in our data, this tone is consistently realized as [55]. Also, all of the level tones have a shortened or "checked" variant (e.g., the level high [5]) that appears in syllables closed by an unreleased stop, as in [wat5] 'twisted'. Following contemporary phonological analysis (M. Chen, 2000; Yue-Hashimoto, 1972), we assume that these variants are the same tonal type but shorter because the syllables that they are realized on are shorter in duration. For the purposes of our analysis, this is actually a conservative assumption because it increases the chance probability of two items having a shared tone (i.e., a 1 in 6 chance versus a 1 in 9 chance), which in turn raises the bar for detecting certain kinds of interactive spreading effects, like what we examine in Section 3.

2.5 Data analysis

Our analysis asks whether, and how often, contextual tone errors occur, while also assessing evidence for interactive spreading effects of tone on the encoding of other linguistic units. In both cases we address these questions by analyzing the frequency of a given speech error pattern. For interactive spreading effects, we are interested in documenting the probability that a particular type of error involves a shared tone. In particular, in Section 3.3 we examine the rates of phonological and lexical substitutions errors when the intended and contextual words do and do not share a tone, and also the effect of phonetic similarity in tone confusions, which involve shared tonal features. In more straightforward cases, we use a chi-square test to test for an association with a shared tone (see Shattuck-Hufnagel (1979) and Stemberger (1989) for illustration and justification of the use of chi-square tests on speech error data). The interactive spreading effect involving phonetic similarity requires a normalization procedure and a correlation analysis that we explain in detail in the Appendix.

3. Results and discussion

3.1 Overview of the data

The distribution of the major error types in SFUSED Cantonese 1.0 is given below in Table 3. This overview serves to introduce the types of speech errors investigated below and also give a sense of the relative frequency of these errors, which is relevant to later discussion.

Table 3. Distribution of error types in SFUSED Cantonese 1.0

Sublexical errors (90.05%)	*N*	Examples
Phonological errors		
Phonological substitution	1,153	mai23 → **b**ai23 ‘rice’
Phonological addition	110	uk55 → **l**uk55 ‘house’
Phonological deletion	90	si22jip22 → si22ji_22 ‘career’
Phonological exchange	3	li55 di55 → **d**i55 **l**i55 ‘these’
Phonological shift	1	tsœt55hœi33 → tsœ**i**t55hœ33 ‘to go out’
Phonological tone substitution	432	hei33kek22 → hei**23**kek22 ‘drama’
Complex set of processes	316	jyn21tsyn21 → jyn21**dz**yn**33** ‘completely’
Other sublexical errors		
Sequential blends	16	lei23 jiu33 → liu23 ‘you must’
Phonetic errors	70	sy55 → s**i-y**55 ‘book’
Morphological errors	26	ba:t33gwa:33geŋ33 → ba:t33gwa:33___ ‘feng shui mirror’
Word and phrase errors (9.95%)		
Lexical substitutions	85	kœi23 ge33 /**jiŋ55man25** ‘his English’ (Intended: ji33da:i22lei22man25 ‘Italian’)
Role mis-selections	14	/**ŋo23** wa:22 ‘I said’ (Intended: kœi23 ‘he’)
Word additions	43	gei25 /**jat55** dyn22 jam55ŋok22 ‘several *one segments of music’
Word deletions	42	lei23 /___ gok33dak55 ‘you think that’ (Intended deleted word: wui23 ‘will’)
Word blends	30	la:m21paŋ21jau23, la:m21jan25 → **la**:**m21paŋ21jan25** ‘boyfriend, man’
Word shifts	9	ham22 /**dzo25** jap22 /∅ ‘fell into’
Phrasal blend	2	hou25 loi22, hou25 do55 lin21 tsin21 → **hou25loi22lin21tsin21** ‘for a long time, many years ago’
Complex set of processes	20	pei33jy21 → bey33 ‘for example’
Total errors	2,462	

A large majority of errors are sublexical errors (over 90%), or errors of word-form retrieval in which the sublexical structure has been mis-selected. These include phonological substitutions, additions, deletions, shifts of segmental information, and also, importantly, tone substitutions, which involve the mis-selection of tone structure. Sequential blends (a.k.a., ‘telescoping’ errors) are likewise sublexical because they merge the form structure of two intended words (Wan & Jaeger, 1998). Below, we discuss how these blends are relevant to determining the nature of tone-to-syllable mapping. The corpus also contains a smaller subset of errors that operate on the level of words and phrases, including lexical substitutions, which we analyze further below. Within both lexical and sublexical errors, there are a large number of speech errors that involve more than one error process, and so are classified as a complex set of processes. Many of the sub-lexical errors exhibiting this complexity combine a tone error with some other error, and so we also examine them in some detail to determine how they may arise.

As a preview of our more detailed analyses below, we answered three basic questions. First, we asked whether there is such a thing as tone encoding at all—that is, whether errors involving tone could accurately be described as existing independently from segmental encoding. With this question analyzed, we then turned to the two key questions discussed in the introduction, asking first whether tone errors are influenced by contextual factors, and then whether tone or segment errors could be unambiguously separated from syllable mis-selections.

3.2 Tone errors: Evidence for the partial independence of tone from segmental encoding

The above overview shows that simple tone substitution errors are the second most common type of speech error.[3] They constitute 17.55% of all errors, and 20.55% of the phonological errors in which a single phonological category is mis-selected. Table 4 breaks down tone substitution errors by type, including complex errors with both a tone and segmental mis-selection. These counts show that, while double tone substitutions and even tone blends do occur, they are exceedingly rare when compared to single tone substitutions. However, complex errors involving both a tone and a segment are more common, and constitute roughly a third of all tone errors.

Table 4. Tone errors by type

Single tone substitutions	419 (62.72%)	miŋ21 → miŋ**22** 'understand'
Double tone substitutions	12 (1.80%)	dzik22dou33 → dzik**55**dou**55** 'until'
Tone blends	1 (0.15%)	gam55lin25, gam25tsiŋ21wan22 → gam**45**= 'this year, relationship luck'
Tone + segmental error	236 (35.33%)	juŋ22 → **dz**uŋ**33** 'use'

The composition of complex tone + segmental errors is broken down further in Table 5. For the most part, the correlating segmental errors mirror the frequencies reported above in Table 3, with phonological substitutions dominating the other patterns.

Table 5. Complex errors with tone and some other segmental error

Substitution	199	jyn21tsyn21 → jyn21**dz**yn**33** 'completely'
Deletion	15	jyn21 → jy_**33** 'finish'
Addition	5	ŋo23 → ŋo**ŋ33** 'I'
Exchange	1	duk22dak22 … foŋ55fa:t33 → duk22d**a**:k**22** … foŋ55f**at22** 'unique … way'
Phonetic error	5	jau23 → **je-au25** 'to have'
Substitution and Deletion	6	goŋ25 gan25 → go_**25** _a**ŋ55** 'talking about'
Substitution and Addition	3	ji23geŋ55 → ji**ŋ33**gi**i**ŋ55 'already'
Deletion and Addition	2	git33gwo25 → gi_**22**gwo**ŋ55** 'result'

When we examined the distribution of simple and complex errors in all phonological errors, we observed that tone errors have a greater likelihood of being complex than segmental errors do. A little over a third of all tone errors are complex, whereas only about one sixth of segmental errors are. Another way of looking at this is to consider just the 316 sub-lexical errors that involve complex processes. Of these, about 75% involve tone errors. This relatively high

[3] The longform of the tone errors discussed below, including their structural attributes, are available from the first author's website following SFUSED > Data Releases.

number of complex errors involving tone, relative to all phonological errors, is confirmed by a chi-square test showing a significant association between complexity and error type ($\chi(1)^2 = 140.19$, $p < 0.001$, with Yates correction), as shown below in Table 6.

Table 6. Simplex vs. complex errors by error type

	Simplex	Complex
Phonological errors	1,804 (85.58%)	304 (14.42%)
Tone errors	432 (64.67%)	236 (35.33%)

The above facts contribute to the question of whether the selection of tone is a separate process from the selection of segments. In particular, the encoding of segments and tone must be separate mappings because the majority of all sub-lexical errors are simple errors that involve one structure without the other. If the selection of tone and segments involved selecting tone and segments together—for example, a complete rhyme with a linked tone—we would not expect such a large number of cases when segments are mis-selected and tones are not, or vice versa.

The observed patterns of sequential blends support this conclusion. Sequential blends involve a wholesale deletion of long strings of segments in a sequence of morphemes or words, for example, *Tennedy* for *Ted Kennedy* (Shattuck-Hufnagel, 1979). As exemplified for Mandarin in Section 1.1., the existence of these errors supports the contention that segments and tone are selected independently because tone can be permuted without the segments associated with it in the intended sequence (Wan & Jaeger, 1998). In practice, these blend errors are difficult to distinguish from a casual speech process that is rather common in Chinese languages called syllable fusion. Syllable fusion can look like a sequential blend error because in a sequence of two successive syllables (which are usually also separate morphemes), rhyme material from the first syllable and the onset of the second can be lost, e.g., /dzi55 dou33/ → dziu53 ‘know’ (W. Y. P. Wong, 2006). To avoid confusing these two phenomena, we examined 37 examples that are ambiguous between the two classes, and excluded 20 cases that had characteristics canonically associated with syllable fusion, including retention of the tones or vowels of both syllables (sometimes producing illicit tones or diphthongs) and deletion of the second syllable’s onset (Cheung, 1986; Kuo, 2010; K. G. Lee, 2003; W. Y. P. Wong, 2006). We also used the talker’s self-corrections as positive evidence for an error (and not fusion), because it is a tacit acknowledgement of a speech error. The remaining 16 cases, illustrated in Table 7, were analyzed as sequential blends, and thus as true speech errors. In ten of these cases, full or partial rhymes are retained with their tones, but five of them (31.25%) involve retention of the tone without the rhyme that it is linked to in the intended word. This again shows that speech errors can explicitly separate tones and segments in phonological encoding, and that these errors occur with some regularity, confirming similar observations found in Mandarin (Wan & Jaeger, 1998).

Table 7. Tone and rhyme retention in sequential blends

Full rhyme	3	dzou22 go33 → gou22 ‘to become a’
Partial rhyme	7	guŋ55dzok33 → dzuk55 ‘work’
No rhyme	5	dou55 dzuŋ22 → dou22 ‘also still’
Other	1	le55 go33 → lə33 ‘this classifier’ (rhyme reduced, so source unclear)

We are careful here to distinguish sequential blends from syllable fusions because of the standard practice of excluding habitual behavior from error data (Dell, 1986). However, this practice does not mean that fusions are not relevant to speech planning involving tone. Indeed,

contemporary accounts of syllable fusion assume that higher-level cognitive processes are at work, including deletion of unstressed syllables and alignment of segments to metrical stress feet (K. G. Lee, 2003; W. Y. P. Wong, 2006). Our data does include several examples of syllable fusions that parallel the data in Table 7, and in particular patterns that exhibit tone permutation without their associated rhymes. These patterns therefore provide additional support for the manipulation of tone apart from segments in these cognitive processes.

As an interim summary, our results support the conclusion that tones and segments can be encoded separately, but also suggest that these processes are not completely independent. That is, if tone encoding were completely encapsulated from the encoding of segments, we would not expect the difference between simple and complex errors shown in Table 6. The incidence of a tone error occurring with a segmental error should have been comparable to the incidence of two segmental errors (e.g., two substitution errors in the same word or a substitution plus a deletion error). We observed instead that mis-selections of tone correlate with mis-selections of segments much more often than mis-selections of two segmental errors (cf. Wan & Jaeger, 1998: 441). Thus, tone and segment encoding must be independent, but only partially so. We discuss possibilities for accounting for the observed interaction between segment and tonal errors in Section 4.3 below by proposing a downstream effect of aligning a selected tone with a process of compiling atonal syllables.

3.3 Tone errors suggest early encoding: Evidence from contextual interactions

Here we discuss what tone errors say about the activation dynamics of tone encoding within a model of speech production. We begin with an analysis of contextual errors, which are speech errors resulting from the surrounding context. In virtually all contemporary models of speech planning, selection of the current word simultaneously involves the activation of elements for both the target word and words in the immediate environment (e.g., Dell (1986) and Stemberger (1982/1985)). The existence of a large number of contextual tone errors, therefore, would support the general claim that tone mis-selection arises from errors in the activation dynamics of tone selection, which would suggest a type of early encoding that coincides with the activation of the surrounding linguistic context.

Table 8 categorizes the observed 419 single tone substitutions into varying types of contextual errors (i.e., perseverations, anticipations, and exchanges) as well as non-contextual errors. Given the high chance probability that the intended tone will be identical to a neighboring tone (approximately a one in six chance, given that there are six tone categories in our analysis), we imposed two distance thresholds in analyzing contextual effects: a standard seven syllable envelope, following Nooteboom (1969), and a more stringent four syllable envelope. Even with the more stringent four syllable envelope, the majority of the tone slips are contextual (76.37%), which compares with the rate of contextual errors for all of our observed speech errors (62.13%). Thus, tone slips are usually contextual, as expected by early encoding.

Table 8. Direction in 419 single tone substitutions, by syllable envelope

Type	7 σ	4 σ	
Perseverations	97 (23.15%)	113 (26.97%)	tsœt55saŋ55 lin21 ^jyt22 ^jat22 /si**22** 'year, month, day, and time of birth' (Intended: si21)
Anticipations	91 (21.72%)	105 (25.06%)	gam25jim23 /dou**33** jan21 ^ge33 'affect other people' (Intended: dou25)
Perseveration + Anticipation	186 (43.39%)	101 (24.10%)	doŋ25bat55^dzy22 ^/hou**22**wan22 'unstoppable luck' (Intended: hou25wan22)
Exchange	1 (0.24%)	1 (0.24%)	dzuŋ22jiu33 → dzuŋ33jiu22 'important'
Non-contextual	44 (10.50)	99 (23.63%)	go33 /ji25saŋ55 dzau22wa:22 'the doctor said' (Intended: ji55saŋ55)

Another form of evidence for a high degree of activation dynamics is the existence of interactive spreading effects. Recall that in the repeated phoneme effect, sounds are more likely to slip if they occur in the same phonetic environment as the source of the intruder sound (Dell, 1984; MacKay, 1970; Wickelgren, 1969). We thus reasoned that, if both segments and tone are encoded relatively early in production, then one would expect a greater-than-chance occurrence of phonological substitutions in which the intended and source syllables share a tone. That is, just as a repeated nearby phoneme can facilitate a substitution, a repeated nearby tone could similarly increase the chance of an error. To investigate this, we aggregated 632 contextual phonological substitutions by the tone of the intended syllable and the tone of the syllable containing the intruder sound, as shown in Table 9. The shaded diagonal gives the frequency of phonological substitutions in which the intended and source syllables share a tone. For example, there were 46 substitutions in which a segment was switched out of a syllable with tone [22], and the intruder sound also came from a syllable that contains [22].

We first examined the relation between intended and source tones using a standard chi-square test, which showed that intended and source tones were not independent, and that table values were not randomly distributed ($\chi(25)^2 = 40.295$, $p = 0.0272$).

Table 9. Intended (rows) and source (column) tones for 632 phonological substitutions

	22	33	55	23	25	21
22	46	18	30	14	22	12
33	25	19	21	11	13	12
55	20	31	48	7	21	18
23	13	5	8	5	10	4
25	34	16	14	11	17	15
21	19	12	21	11	15	14

Given this result, we then considered the hypothesis that this lack of independence was due to a "repeated tone effect", i.e., higher than expected values on the diagonal in Table 9. In Table 10, we show the ratio of observed/expected values, given the error data in Table 9, to get a sense of the frequency of the repeated tone pattern relative to the error corpus. Here, we see that all of the values on the diagonal are greater than one, including cases like [55]/[55] with rather high values, suggesting an over-representation of phonological substitutions when the intended and source words share a tone.

Table 10. Observed/expected for intended and source tone in phonological substitutions

	22	33	55	23	25	21
22	1.30	0.79	0.94	1.06	1.00	0.71
33	1.00	1.18	0.93	1.17	0.83	1.00
55	0.56	1.34	1.47	0.52	0.93	1.05
23	1.16	0.70	0.79	1.19	1.43	0.75
25	1.28	0.94	0.58	1.10	1.02	1.18
21	0.83	0.82	1.02	1.28	1.05	1.28

These facts, however, do not factor in the baseline frequencies in a way that gives us a measure of statistical significance. To do this, we conducted individual goodness-of-fit tests on the rows in Table 9 in which expected values are scaled to the known token frequencies for Cantonese tones (Leung, Law, & Fung, 2004). In particular, we tailored our expected counts to match token frequency in order to remove frequency as a confound. To account for multiple test of null hypotheses, we also use Bonferroni correction of the chi-square goodness of fit tests. In particular, we multiply all *p* values by 6, the number of hypotheses, to confirm the result. The results given in Table 11 show that there is indeed a repeated tone effect, but it seems to interact with the tone type because we only find significant effects with the low level [22] and high level [55] tones. The same tests were conducted using type frequency instead of token frequency, with the same outcomes. Thus, there is clearly a repeated tone effect, but it seems to be limited to two of the six tones.

Table 11. Goodness of fit tests for repeated tone effect

	Observed		Expected				
Tone	Shared	Not shared	Shared	Not shared	$\chi(1)^2$	*p*	Bonferroni correction
22	46	96	29.84	112.16	10.41	0.0013*	0.0075*
33	19	82	19.32	81.68	0.002	0.9643	> .99
55	48	97	30.97	114.03	11.219	0.0008*	0.0049*
23	5	40	4.81	40.19	0.022	0.8821	> .99
25	17	90	17.70	89.3	0.003	0.9563	> .99
21	14	78	10.36	81.64	1.072	0.3005	> .99

Another type of interactive spreading effect that we considered is the similarity of intended and intruder tones. It is well-known that segmental errors are affected by similarity: The intended sounds that are supplanted in errors tend to be replaced by similar sounds, at least in English and German where this question has been investigated (MacKay, 1970; Shattuck-Hufnagel & Klatt, 1979). Table 12 shows the confusability of tone in the 419 single tone substitutions in our corpus, i.e., the counts of intended tones (rows) being replaced by intruder tones (columns). A chi-square test on the row and column totals indicates that the six tones do not differ as an intended or intruder tone ($\chi(5)^2 = 3.046$, $p = 0.6929$), so the tones are largely

symmetric as inputs and outputs of errors. However, it appears that similar tones tend to be substituted more often than dissimilar tones. For example, there are 70 substitutions of [22] and [33] (in both directions), as opposed to only 13 substitutions of [22] and [55], which are phonetically more distinct than [22] and [33].

To assess this similarity effect systematically, we required a measure of tone similarity, as well as a test for correlations between confusability and similarity. In the Cantonese tone literature, there is no widely agreed upon procedure for assessing the phonological similarity of tones because there is no consensus on the critical tone features (Barrie, 2007; J. L. Lee, 2012). We therefore used an estimate of phonetic distance between citation tones, calculating similarity as the inversion of this distance. Here we assumed that phonetic distance for comparisons involving contour tones could be estimated by the sum of the distance between the onsets and offsets of intended and intruder tones in their phonological description. For example, [23] and [21] have a distance of 2 because the distance in their tonal onsets is 0, but the distance in their offsets is 2 (one tone ends at level 3, while the other at level 1). For level tones, we assume that the distinction between onset and offset is collapsed in comparisons between two tones, because they are generally regarded as a single tone.

Table 12. Confusion matrix for only single tone substitutions indicating intended tones (rows) and intruder tones (columns)

	22	33	55	23	25	21
22		37	7	25	18	26
33	33		7	16	16	6
55	6	17		0	13	2
23	16	9	7		18	11
25	20	20	20	15		1
21	32	5	2	14	0	

We use a Mantel test (Glerean, 2014) to test for a correlation between the similarity matrix, which contained our custom measures of phonological distance for Cantonese tone, and the confusion matrix for tone errors in single tone substitutions shown above. Prior to applying this test, however, we also removed any effects of the overall frequency of a particular tone error by normalizing the data in the confusion matrix (see Appendix). Results showed a positive correlation, $r = 0.562$, $p = 0.0437$ (simulated p value with 5,000 permutations). In short, tone slips appeared to mirror segmental slips in that confusability was affected by similarity of tone.[4]

[4] There are a set of sound changes in progress in Hong Kong Cantonese that could have affected this result because they involve mergers of pairs of similar tones: 23/25, 22/33, and 22/21 (see Bauer et al. (2003); Mok et al. (2013)), which could have been mis-heard by our collectors as errors. Two of our data analysts re-examined the data with this issue in mind, and after excluding a small number of cases that could have been due to the mergers, they were confident that the remaining cases are indeed tone slips. This view is supported by an analysis of speech errors corrected by the speakers themselves: most of our speakers produced an erroneous tone, and then explicitly corrected that tone to one that participated in at least one of these mergers, reflecting the fact that merger pairs are indeed distinct tones in the minds of our speakers. Furthermore, if our tone slips are really due to tone mergers, we would expect far higher numbers of slips than actually observed. This is because five of the six tones participate in mergers, and so, if mergers were truly misheard as errors, there would have been an opportunity for such a mistake in almost every word.

The above two effects document interactive spreading among form elements traditionally activated in phonological encoding. However, such effects are well-documented in speech error studies between grammatical and phonological encoding, including between lemma selection and phonological encoding of segments (Dell et al., 1997; Fay & Cutler, 1977; Fromkin, 1971; Goldrick, Folk, & Rapp, 2010). We have already discussed the importance of Wan and Jaeger's (1998) finding that lexical substitutions in Mandarin have a greater than chance probability of sharing a tone. Repeating such a finding in Cantonese is thus relevant to the question of how tone is encoded, because this result demonstrates interaction between tone encoding and lemma selection, which is standardly assumed to be initiated prior to phonological encoding (Bock & Levelt, 1994; Dell, 1986; Levelt et al., 1999).

Unfortunately, it is not possible to rigorously investigate such a pattern in our corpus because of insufficient data. When role mis-selections and word blends are excluded, our corpus only contains 85 straightforward lexical substitutions, and 40 of these are in polysyllabic words that are hard to interpret (because more than one tone is selected, making accidental convergence higher). Of the 45 lexical substitutions involving monosyllabic words, as in /**wui23** siu25sam55di55/ 'will be more careful' (Intended: jiu33 'need to be'), 13 of them (roughly 29%) share a tone in the intended and error words. We cannot do the same tests of independence and goodness-of-fit that we did for individual tones in phonological substitutions (Table 9 and Table 11) because of small cell counts. However, such a test on shared versus not-shared tone (Table 13), based on a one-in-six expected frequency (as used by Wan and Jaeger (1998) for Mandarin lexical substitutions) is significant ($\chi(1) = 4.84$, $p = 0.0278$), suggesting that lexical substitutions with shared tones are in fact over-represented. Given our inability to factor in baseline frequencies, this result can only be taken as suggestive, but one aspect of lexical substitutions parallels the facts of tone in phonological substitutions. Six of 11 lexical substitutions with the low level tone [22] in the error word also had [22] in the intended word, and three of seven had the same pattern for the high level tone [55]. These are the same shared tones that were over-represented in the phonological substitutions in Table 9, suggesting that the interactive encoding of lemmas in lexical selection is also associated with specific tones.

Table 13. Shared vs. not shared tones in monosyllabic lexical substitutions; Expected = 1/6 probability

	Shared	Not Shared
Expected	7.5	37.50
Observed	13	32

As an interim summary, there seems to be strong evidence for early encoding of tone because of its interactive nature. The majority of tone slips result from an interactive process because they perseverate or anticipate a nearby tone. In addition, there is more subtle evidence for interactivity because shared tones tend to increase the incidence of other kinds of errors, like phonological substitutions and lexical substitutions, though our evidence for the latter is only suggestive. Finally, phonetically similar tones slip more than dis-similar tones, suggesting that shared features lead to higher error rates, another hallmark of interactive phonological encoding.

3.4 The status of segments versus syllables in speech encoding

Another major debate raised in the introduction concerns the status of segments and syllables in phonological encoding and the predictions of the proximate unit hypothesis (O'Seaghdha et al., 2010). As discussed in the introduction, the proximate unit hypothesis posits

a language-specific type of activation that occurs directly after lemmas are chosen in lexical selection. Languages like Mandarin Chinese, largely on the basis of priming studies, have been argued to have the syllable as the proximate unit that is activated after the lemma level. In particular, atonal syllables, i.e., the consonant and vowel string of a single syllable without specifying a tone, are argued to be a privileged unit in production. Notably, segments are secondary in that they are selected after atonal syllables in Mandarin (Figure 2). However, this literature has not been extensively explored by examining naturalistic speech production. Here, we ask if this theory is viable in different forms in Cantonese, given the evidence reviewed above for both syllable and segment-level priming effects.

In general, speech errors involving entire syllables are exceedingly rare in languages like English and Dutch (Nooteboom, 1969; Stemberger, 1983). Against this backdrop, J.-Y. Chen (2000) investigated the frequency of whole syllable errors in a corpus of Mandarin errors and found some evidence for a privileged status of syllables in encoding. In particular, ten (8.4%) out of 119 sound errors in this corpus were argued to involve whole syllables that could not be reanalyzed in other ways. Furthermore, it was demonstrated that this 8.4% rate of occurrence is greater than chance if chance is calculated as the probability of independent errors involving the component sub-syllable sounds. One concern with this result, however, is the rather low number of sound errors in this corpus, which clearly factors into the chance estimates of independent segmental errors. For comparison, the Stemberger corpus has 6,300 speech errors in English, of which 3,660 are sound errors (Stemberger, 1983). This latter corpus is reported to have 13 whole syllable errors, which is 0.36% of all sound errors, far lower than Chen's estimate for Mandarin. While it could be the case that Mandarin and English simply have different rates of syllable errors, recall that speech error data collection is plagued with methodological problems (Alderete & Davies, 2018; Ferber, 1995), and so it seems more prudent to first understand why the Chen corpus (2000) has such a low rate of sound errors.

Another way to investigate the role of segments relative to syllables in speech errors is to examine unambiguous speech errors. Whole syllable errors are ambiguous because they can be analyzed as either the mis-selection of an entire syllable or the mis-selection of the sounds contained in that syllable. Many speech errors involving individual segments, or strings of segments like VC rhymes, are likewise ambiguous because the replacement by the intruder segments is still consistent with a syllable level analysis if the intruder sounds are replaced with the other unchanged segments. For example, the apparently segmental error of /*t/ →/k/* in *top → cop* can be analyzed as either a segmental substitution, or the wholesale replacement of the syllable [tɑp] for [kɑp]. Indeed, Mandarin has a far smaller inventory of possible syllables, making the listing and productive use of syllables a tractable problem (T.-M. Chen, Dell, & Chen, 2004), one which increases the ambiguity between syllable and segment errors. There are, nevertheless, certain kinds of errors that are unambiguously segmental for the simple fact that they are impossible syllables. While sound errors tend to obey phonotactic rules (Boomer & Laver, 1968; Wells, 1951), meaning that sounds generally slip into well-formed syllables, recent evidence has shown that this constraint is far weaker than previously assumed (Alderete & Tupper, 2018).

We have investigated the ambiguous and unambiguous nature of sound errors in SFUSED Cantonese with these issues in mind. Table 14 gives the frequencies of 1,357 sound errors and breaks them down into classes that are ambiguous between segmental or syllabic mis-selections, and those that are unambiguously segmental errors. In all major error types, ambiguous syllable/segment errors like *guk → kuk* dominate the data, accounting for roughly

85% of all errors. There are also a non-trivial number of ambiguous whole syllable errors in substitutions and additions, e.g., *kek → haŋ*, that could be either syllable level errors or combined segmental errors.

Crucial to our discussion, however, is the relatively large number of *unambiguous* sound errors. There are 147 examples in our corpus that violate the principles of well-formed syllables in Cantonese (Bauer & Benedict, 1997). For example, the substitution *sa*[t] → *sa*[s] contains a syllable final [s], which is not possible in Cantonese. Likewise, the deletion of syllable final [p] in *sa[p]* → *sa* results in the short vowel [a] in an open syllable, which is again outlawed by Cantonese phonotactics. The frequency of errors with phonotactic violations is 10.8%, which is a bit higher than what has been reported for English (Alderete & Tupper, 2018). This may be due to the relatively high number of non-native sounds in the Cantonese corpus, which accounts for roughly half of these cases. But even if these are excluded, the remaining 5% of the data are errors that unambiguously involve segments and not syllables.

Table 14. Ambiguous and unambiguous sound errors

Pattern	Example[5]	Count
Substitution (*n*=1,159)		
Ambiguous whole syllable	hei33[kek]22 → hei33**haŋ**22	58 (5%)
Ambiguous syllable/segment	kei21[g]uk33 → kei21**k**uk33	980 (84.56%)
Unambiguous segment	kei21sa[t]22 → kei21sa**s**22	121 (10.44%)
Addition (*n*=110)		
Ambiguous whole syllable	hai22mai22 → hai22**tsə22**mai22	4 (3.64%)
Ambiguous syllable/segment	ji21 → ji**t**21	86 (78.18%)
Unambiguous segment	mou21la:55la:55 → mou**t**21la:55la:55	20 (18.18%)
Deletion (*n*=88)		
Ambiguous whole syllable	N/A	0
Ambiguous syllable/segment	si22ji[p]22 → si22ji22	82 (93.18%)
Unambiguous segment	luk22sa[p]22 → luk22sa22	6 (6.82%)

These results contribute to the proximate unit debate in two ways. First, it provides new counts of errors that may involve whole syllables (and are possibly not simple slips of segments). These examples account for approximately 4.57% of the sound errors examined above, which is far below the rate documented in J.-Y. Chen (2000) for Mandarin, but also much more common than the rate reported in Stemberger (1983) for English. While conjectural, we believe these facts suggest a middle ground between the two empirical patterns discussed above, and offer some support for the idea that syllables can be mis-selected as wholes in speech errors. However, given that many factors contribute to the actual frequencies, and some production models are not specifically designed to predict speech errors (Levelt et al., 1999), the problem of predicting the precise rates of syllable errors will have to be taken up in future work.

[5] The English glosses for the intended words, ordered top to bottom, are as follows: ‘drama’, ‘chess game’, ‘actually’, ‘is it or is it not?’, ‘and’, ‘for no reason’, ‘career’, ‘sixty’.

In addition to this conclusion, our data also strongly suggest a role for a mechanism that selects individual segments, separate from syllables. The abundance of sound errors resulting in illicit syllables requires this. We believe this teases apart two versions of theories that give syllables special prominence in phonological encoding that have not yet been carefully examined: (i) a strong form in which syllables are selected immediately after lemmas, which leads directly to the activation of the syllable motor programs, or (ii) a weaker form of the hypothesis that selects syllables first, then the component segments, and then syllable motor programs. It should be said that all syllable-based theories that we are aware of are the latter type and include a mechanism for selecting segments, including recent versions of WEAVER++ tailored to Chinese languages (Roelofs, 2015) and all theories of phonological encoding that assume the proximate unit hypothesis (J.-Y. Chen et al., 2002; J.-Y. Chen et al., 2016; O'Seaghdha et al., 2010); see Figures 1 and 2 for a visualization of these encoding mechanisms. A comparison of these two versions, though somewhat rhetorical, gives us an empirical basis for including a selection mechanism for segments. The strong form of the proximate unit hypothesis simply does not allow for illicit syllables because there is no step in which the selection of segments can go awry.

Returning to the question of encoding tone, the above findings lead to a new empirical question that could potentially support our conclusion from Section 3.3, that word-form encoding in Chinese languages requires a separate mechanism for tone. Given the assumption of the syllable as the proximate unit, we can look for a parallel in tone errors to the unambiguous segmental errors examined above. In particular, are there tone errors that result in tonal syllables that are otherwise outlawed in the language? The existence of illicit tone + syllable combinations in tonal errors would constitute evidence similar to the sequential blend facts discussed in Section 3.3: Tones must slip independently of the segments and syllables that are associated with them, because illicit tone + syllable combinations are not stored under any proposed theory.

There are lots of possibilities for documenting illicit tone + syllable combinations, but some of them, like combinations of laryngeal settings of onset consonants relative to tone (see Yue-Hashimoto, 1972: 110 ff.), are not universally accepted as constraints on tone structure. However, one generally agreed upon restriction on tone is that so-called checked syllables, or syllables that end in unreleased stops /p t k/, are restricted to shortened versions of level tones. That is, these syllables do not combine with any of the three contour tones [25], [23], [21] (Bauer & Benedict, 1997; Yue-Hashimoto, 1972). We examined our corpus with this restriction in mind, and indeed found nine examples in which a tone slip resulted in an ungrammatical checked syllable + contour tone combination, as in /fa:t33jin22/ → **fa**:**t35**jin22 ‘to discover’. These nine examples constitute 3.5% of the 257 cases that could result in such an outcome, a non-negligible number given that checked syllables are under-represented generally in Cantonese (Leung et al., 2004). Therefore, like unambiguous segmental errors, it also appears that there are tone slips that result in phonological illicit combinations. This finding provides additional supporting evidence for the conclusion established in Section 3.3, that phonological encoding requires an independent mechanism for encoding and selecting tone.

4. Discussion

Our analysis of tone errors in Cantonese provides insights into the cognitive mechanisms of tone production and planning, and our results broadly share some similarities, but also some

differences, with prior studies of tone errors. Here, we discuss how our results compare to prior work in order to offer a more cross-linguistically robust empirical picture.

Our work broadly supports two principal generalizations about Cantonese tone production and planning. First, tone and segmental encoding appear to be partly independent of each other insofar as one can occur without the other. Yet this observation is qualified by the observation that tone errors are more likely to occur together with segmental errors when producing a syllable, than when segmental errors occur without tone errors. Second, Cantonese tone errors appear to be comparable to other types of phonological errors in speech production, which suggests an early and interactive encoding process. The last section of this discussion integrates these observations with existing models of speech planning and production in Chinese languages.

4.1 Early versus late encoding of tone and the frequency of tone errors

A fundamental debate in the literature is whether tone errors occur with a frequency comparable to other phonological errors (Gandour, 1977; Shen, 1993; Wan & Jaeger, 1998), or whether tone errors are negligible, like stress errors (J.-Y. Chen, 1999; Kember et al., 2015). An overview of the descriptive statistics of tone errors across studies (Table 15) clearly supports the former view. Three studies in three different languages (Mandarin, Cantonese, and Thai) show that tone errors are relatively common. Furthermore, when counted as a percentage of sound errors, all studies report a relatively high percentage of tone errors, including J.-Y. Chen (1999), who argued tone errors are exceedingly rare. It turns out that Chen's dataset contains a comparatively low number of sound errors (16.2% of all errors), and while some of the reported 24 ambiguous tone errors may have an alternative analysis, it seems clear that even the tone errors in this study are well-represented as a percentage of sound errors. Together, these studies strongly suggest that tone errors are non-negligible, comprising around 13%-20% of all sound errors in three large-scale studies.

Table 15. Quantitative summary of major tone error studies

	Wan & Jaeger 1998	Chen 1999	SFUSED Cantonese 1.0	Gandour 1977
Language	Mandarin	Mandarin	Cantonese	Thai
All errors	788	987	2,462	*unknown*
Sound errors	597	160	2,105	*unknown*
Tone errors	78	24	432	350
Tone % of sound	13.07%	15%	20.52%	*unknown*

Moreover, our analysis of these tone errors suggested that error types are highly influenced by the surrounding context, which indicates where tone is encoded in the activation dynamics of phonological encoding. Nearly all models assume that if a phonological unit is actively selected in phonological encoding, then one would expect large numbers of contextual errors associated with that unit. Our work therefore suggests that tone is encoded early, and is moreover subject to interactive spreading effects that are hallmarks of early encoding. Supporting our position is data from Wan and Jaeger (1998), who found an interactive spreading effect involving shared tones in lexical selection errors, which is similar to what we found in lexical errors, and in our analysis of phonological substitution errors. Moreover, we also found a similarity effect in a sub-analysis of single-tone substitutions, which suggested that, just like for segments, intended and substituted tones have a tendency to be more similar to each other. In

sum, our results point to a common level of processing for tone and segments in speech production, which is consistent with the idea that tone is encoded early in planning.

One might object to this line of argumentation by observing that tone errors are less common than other segmental errors, like consonantal slips, under any analysis. This observation is consistent with the findings of experimental paradigms designed to elicit tone errors. For example, Kember et al. (2015) used a tongue twister experimental paradigm and observed that, while tone slips can be induced through priming at a non-negligible rate, they nonetheless occur less frequently than segmental errors. Our data also show this pattern, as shown in Table 16, which illustrates the relative frequencies of single consonant, vowel, and tone substitutions in our study.

Nevertheless, our view is that many factors influence the rates of error frequency, and precise predictions of error frequency need to attend to all of these factors. For example, the high rate of consonantal errors may relate to the fact that words more often begin with consonants than vowels and tones, and so they are more prone to the word-onset effect that has been documented for English and German (MacKay, 1970; Shattuck-Hufnagel, 1987; Wilshire, 1998). It is important to point out, however, that the word onset effect is not found in all languages (Abd-El-Jawad & Abu-Salim, 1987; Berg & Abd-El-Jawad, 1996; Pérez et al., 2007), and it is unknown if word-onsets are especially prone to error in Cantonese or Mandarin. Another fact is that in Cantonese, and Mandarin as well, consonants can fill both onset and coda positions, so there is potential for mis-selection in two places per syllable. This contrasts with vowels and tone, which are selected only once per syllable. Finally, it seems highly likely that error frequency relates to inventory size and the nature of the selection mechanism within that inventory. As shown in Table 16, filling a consonant onset position in Cantonese involves selecting one consonant out of 19, and so the potential for mis-selection of similar sounds is greater than with vowels and tones (inventory size counts are from Bauer and Benedict (1997)). A related idea is that tonal differences may be simpler to articulate. Indeed, the four-way tonal contrast in Mandarin is analyzed with just two gestures in Gao (2009), which may require less gestural complexity and coordination than the gestural analyses of segments. If true, then articulatory complexity could explain some of the disparities here. In sum, we argue that asymmetries in error types, or simply counting the number of segmental versus tone errors, can tell us very little about levels of speech encoding. Rather, these asymmetries are likely a consequence of several selection factors, such as those discussed above.

Table 16. Single item substitutions and inventory size in Cantonese phonology

Unit	Error frequency	% of total	Inventory size
C	714	49.69%	19
V	304	21.15%	10
Tone	419	29.16%	6

Another piece of evidence related to the level of encoding is the interaction of tone slips with tone sandhi rules, which change surface tones in the vicinity of a sandhi trigger tone. For example, Mandarin [21] [21] →**[35]** [21] tone sandhi is generally assumed to be late in phonological processing. Tone sandhi is not attested in Cantonese, and so our data cannot speak to this phenomenon, but evidence from Wan and Jaeger's (1998) analysis of Mandarin tone errors found strong evidence that sandhi takes place after the encoding of tones. As an illustration from that study (p. 444), consider: /na35 [jow21 maj51] pɑw51-tʂi21/ → na35 **jow35**

maj21 pɑw51-tʂɨ21 (‘Where is the place selling newspapers?’). Here, the tone of the morpheme *maj* slips from [51] to [21], which in turn provides the trigger tone for the [21] to [35] tone sandhi in the preceding syllable. If the tone slip happens first in phonological encoding, it provides the correct input for tone sandhi at a later level of phonetic processing. It is not clear, however, how this interaction is explained if tone is not actively selected in phonological encoding (i.e., represented with diacritic labels and spelled out later). In Chen’s (1999) model of late tone encoding, for example, tone sandhi is actually ordered prior to the phonetic spell-out of tone, which is inconsistent with these facts.

4.2 Evidence for a (partially) separate mapping of tone and segments

If it is true that tone is selected in phonological encoding, is this selection part of the same process that selects segments, or is it a separate mapping? Our results are consistent in part with the findings of Wan and Jaeger’s (1998) analysis of Mandarin: Errors involving single segments without tone and single tones without segments constitute the vast majority of all phonological errors. If phonological encoding involved a selection of both segments and tone together—for example, selecting from syllable rhymes in Cantonese pre-specified for specific tones—such patterns of single errors should be more the exception rather than the rule. Furthermore, sequential blends in both Mandarin and Cantonese support this view because tone structure can be permuted to a rhyme that it is not associated with in the intended word. Like these blends, ungrammatical tone + syllable combinations also support an independent tone selection mechanism, because these illicit combinations are not stored. Moreover, other evidence in the field—particularly studies that have used priming methods to investigate speech production—have suggested the existence of planning units where segmental content is specified independently of tonal information. That is, native speakers of tone languages are able to represent ‘atonal’ syllables in their speech planning, as well as ‘tonal’ syllables (J.-Y. Chen et al., 2002; O'Seaghdha et al., 2010). Together, this evidence suggests that tone selection is distinct from segmental selection in phonological encoding.

At the same time, our data suggest that the production of one type of speech unit is not entirely encapsulated from the other. As documented in Section 3.2, complex errors (i.e., involving tone and some other segmental error) are roughly twice as common as complex errors involving two segmental changes. That is, tone errors correlate with segmental errors more often than would be expected, compared to other complex errors. Importantly, these errors are not accounted for by simply assuming some or all tones are selected together with segments within some planning unit. This is because, in many of these complex tone + segment errors, the tones and segments slip independently, and they can produce combinations that are not valid words.

Broadly, our data suggest the need to reconsider the role of tone in speech production and planning, given simultaneous evidence for independent and linked phonological encoding of segments and tones. In the final section below, we offer an account that integrates across the cumulative evidence from speech errors in tone languages, as well as other studies of tone language production.

4.3 Model considerations

Following from an analysis of errors in naturally produced speech, while also considering evidence from form preparation studies (J.-Y. Chen et al., 2002; O'Seaghdha et al., 2010) and elicited error paradigms (Kember et al., 2015), we propose a revised version of O’Seaghdha’s planning model (Figure 2), which integrates the novel findings discussed here into prior speech

planning frameworks proposed for Chinese. Our revisions, shown in Figure 3, attempt to account for four key aspects: i) an early encoding of tone with spreading activation that occurs at roughly the same time as segmental encoding, ii) distinct stages for tone and segment selections, iii) distinct stages of syllable and segment selections, and iv) downstream interactions between the mapping of selected tones and segments that can account for the large number of tone + segment errors.

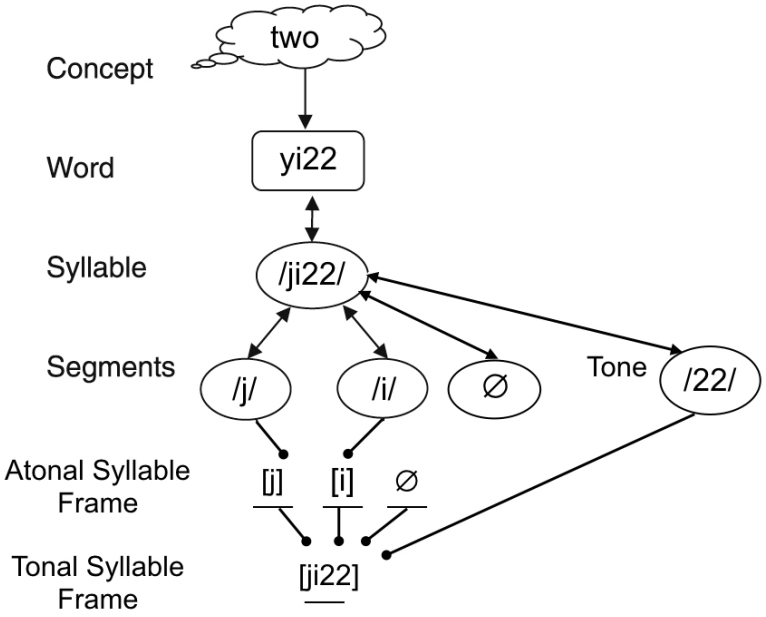

Figure 3. Tone encoding with tone selection and tone-to-atonal syllable mapping

Before discussing our theoretical results, it should be made clear that the ideas behind the proximate unit hypothesis were developed to account for form preparation effects in implicit priming studies (see Section 1.2), and much less attention has been paid to how specifically speech errors arise. Spreading activation models (Dell 1986, et seq.) have rather different accounts of speech errors than models designed for form preparation like WEAVER++ (Levelt et al., 1999). Speech errors are produced in spreading activation models by insertion rules that mis-select items from a larger network of comparable items, because the context for the retrieval of that item (and noise) leads to higher overall activation than the intended item. While the original proximate unit hypothesis does not directly espouse these insertion rules, and therefore does not necessarily see errors as mis-selections, this is how we conceive of the mappings from the Syllable to Segments levels and Syllable to Tone levels. In particular, tone is actively selected from an inventory of tone form elements, and this inventory is within a larger network of form elements that includes syllables and segments.

Our revised model maintains several of the hallmark characteristics of O'Seaghdha et al.'s (2010) proposal, including the proximate status of the syllable. However, we have modified the original proposal in two major ways. First, we include a tone selection process that occurs between the Syllable and Tone levels, which is equivalent to selection between the Syllable and Segments levels. This modification explains the prevalence of tone errors and interactive spreading effects on them arising from their interaction with lexical selection and phonological encoding. The bi-directional arrows indicate backward and forward spreading of an activation

signal from tones to syllables, and then onto segments and words. These activation flows account for the observed interaction between tone and segments (via tone-syllable-segments) and tone and lemmas (tone-syllables-words). Because of the lack of consensus on the phonological features of tone (see Section 3.3), we do not show tone features, but bi-directional spreading of tone to features is assumed to account for the phonological similarity effect in tone selection, like it does for segments (Dell, 1986). Moreover, positing two selection steps allows for the possibility that certain experimental priming methods may isolate activation of units more proximate to the lexicon (i.e., O'Seaghdha et al., 2010; Chen & Chen, 2013, etc.), while still allowing for other methods to show effects at the level of individual segments and tone (i.e., a picture-word interference task, as well as our current speech error analysis).

Second, we also envision a two-stage mapping after segments and tones have been selected, such that an 'atonal' syllable frame is first assigned based on the segments, followed by a secondary and later mapping of tone onto the syllable frame after it has been filled with segments. This assumption gives a straightforward account of the difference between complex segmental and complex tone errors documented in Section 3.2. In this model, there are two mapping steps after selecting the target segments, once to an atonal syllable, and then again to a tonal syllable, i.e., the atonal syllable linked with the correctly selected tone. Because the tone to atonal syllable mapping follows the creation of the atonal syllable itself, a segmental error changes the talker's expectations about the tonal syllable, which can in turn lead to new errors in this mapping. For example, the intention to produce lemma /yi22/ leads to the atonal syllable [[j] [i] [Ø]], which the talker expects to align with the correct tone [22]. A mis-selection prior to the creation of [[j] [i] [Ø]], for example, selecting a different vowel, like [[j] [y] [Ø]], can remove the associations a talker may use in tone-to-atonal syllable alignment, therefore leading to a tone error as well. The result of this architecture would be a higher number of tone + segment errors than tone errors alone, consistent with our findings. However, double segment errors will not arise in this way because all segments are associated with the syllable frame in the same mapping step, thus accounting for the fact that these latter errors are much less frequent. The associations between selected segments and complete syllables crucial to deriving this result can be learned with existing learning architectures, like the multi-layer network developed in Warker and Dell (2006) for precisely this kind of mapping in phonotactic systems. As far as the relative timing of segments and tone, we are not aware of any evidence that would conclusively distinguish the two structures. However, evidence from event related potentials suggests that information about segments and tone are accessed concurrently and in parallel in implicit priming (Zhang & Zhu, 2011), which is also consistent with our model in Figure 3.

The empirical results of this work strongly favor a selection mechanism for tone that is on a par with segments. Proposing such a mechanism obviates the need to map tone to structural frames, e.g., the Tonal Frames of Roelofs (2015). This in turn leads to the question of how tonal syllables are ordered serially in polysyllabic words. Because our focus is on the facts and analysis of encoding tone, we do not argue for a particular serial order mechanism, but one idea based on the nature of prosody in Chinese languages has strong potential. To begin, tonal syllables could be ordered directly in a sequential network in which a plan is associated with a sequence of syllables, similar to the way Dell, Juliano, and Govindjee (1993) have associated a plan with a sequence of segments. Because sequential networks such as these are less common than non-sequential networks in which multiple units are simultaneously activated (a class that includes all models discussed above), it will be prudent to sketch a possibility within this mainstream view. In a non-sequential approach, phonological encoding in Chinese languages

could employ word-shape frames based on metrical structure rather than tonal structure. One might object to this approach on the grounds that Cantonese is a tone language and metrical structure is not appropriate linguistic analyses. However, this objection is based on a common misconception about tonal languages, namely the existence of tone precludes prosodic structure (see Poser (1984) and Yip (2002) for discussion). The phonology of Chinese languages provides abundant evidence for prosodic structure from a wide range of facts, including phonotactics, tone alignment, tone sandhi, compounds, and importantly phonetic stress (Duanmu, 1995, 2007; Selkirk & Shen, 1990; Shih, 1986). This evidence is so pervasive that it is difficult to imagine analyses of tone in Chinese languages without prosodic feet. The fact that this evidence has equal importance in Cantonese (W. Y. P. Wong, 2006; Yip, 1992) strongly suggests that prosodic feet could be used for building structural frames in this language. Such frames, in turn, can be used to order syllables in a way parallel to how it is used in English and Dutch (Levelt et al., 1999), by positing a prosodic frame that is associated with selected syllables (e.g., effectively replacing Roelofs tonal frames with prosodic frames in Figure 1). Identifying the correct serial order mechanism for polysyllabic words will have to be investigated in future work, but structural frames based on prosodic feet seem like the best place to start given the facts of Chinese languages.

The evidence that we observed from Cantonese tone errors supports, at a minimum, a two-stage architecture that incorporates both the early phonological encoding of tone and a later mapping of tone (along with other metrical information), which is susceptible to complex interactions with segmental mapping. Future work must verify some of the predictions of this model, perhaps using experimental methods or investigating additional tonal languages. For example, it is currently unclear whether "atonal" representations can be experimentally isolated in Cantonese speech production studies, and it is also unclear how precisely metrical information at the phrasal level (e.g., sandhi processes in other Chinese languages, as well as intonational information) would factor into the "tonal" syllable frame. Nevertheless, we are hopeful that this novel data about Cantonese speech errors provides a new theoretical platform from which hypothesis-driven studies can advance our understanding of the cognitive processes underlying the production and planning of tone.

Appendix

How to investigate the speech error examples in SFUSED Cantonese 1.0

The examples from the above tables can be explored further in the associated database by searching on the following record ID values. The raw data is available from the first author's webpage. Table 1: 1125, 744; Table 3: 30, 338, 192, 2918, 2394, 758, 17, 312, 777, 591, 1166, 715, 1045, 369, 151, 460, 1437, 2543; Table 4: 430, 111, 186, 395; Table 5: 17, 1020, 3492, 2622, 266, 2043, 2609, 3607; Table 7: 3013, 3714, 2593, 1061; Table 8: 210, 417, 337, 1964, 1300; Table 14: 764, 998, 1892, 456, 842, 1209, 192, 2807.

Mantel test for tone confusion matrix and similarity

For the test of correlation between similarity and tone confusability, the matrix in Table 12 was first normalized by row totals to remove the effect of frequency. All entries off the diagonal were then inverted because we seek to correlate confusability with similarity, but phonetic distance is the inversion of similarity. The matrix was then symmetrized because the intruder-intended order

has no impact on similarity. Applying the bramila_mantel function in Matlab (Glerean, 2014) to this matrix and the symmetrized matrix for phonetic distance produced both a correlation coefficient and a *p*-value for significance testing.

Acknowledgements

We are grateful to Paul Tupper and Stephen Matthews for valuable comments and assistance, and Gloria Fan, Tsz Ying Ng, and Macarius Chan for their tireless efforts in data collection and classification. This work is supported in part by an insight grant from the Social Sciences and Humanities Research Council of Canada (435-2014-0452).

References

Abd-El-Jawad, H., & Abu-Salim, I. (1987). Slips of tongue in Arabic and their theoretical implications. *Languages Sciences, 9*, 145-171.

Alderete, J., & Chan, Q. (2018). *Simon Fraser University Speech Error Database - Cantonese (SFUSED Cantonese 1.0)*. Burnaby, British Columbia, Canada.

Alderete, J., & Davies, M. (2018). Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings. *Language and Speech, DOI: 10.1177/0023830918765012*.

Alderete, J., & Tupper, P. (2018). Phonological regularity, perceptual biases, and the role of grammar in speech error analysis. *WIREs Cognitive Science, 9*, e1466. doi:10.1002/wcs.1466

Barrie, M. (2007). Contour tones and contrast in Chinese languages. *Journal of East Asian Linguistics, 16*, 337-362.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology* (Vol. 102). Berlin: Mouton de Gruyter.

Bauer, R. S., Cheung, K.-h., & Cheung, P.-m. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change, 15*, 211-225.

Berg, T. (1988). *Die Abbildung des Sprachproduktionsprozesses in einem Aktivationsflußmodell [The illustration of the speech production process in an activation flow model]*. Tübingen: Max Niemeyer Verlag.

Berg, T., & Abd-El-Jawad, H. (1996). The unfolding of suprasegmental representations: A cross-linguistic perspective. *Journal of Linguistics, 32*, 291-324.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin and Review, 3*, 395-421.

Bock, K., & Levelt, W. J. M. (1994). Language production. Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945-984). San Diego: Academic Press.

Boomer, D. S., & Laver, J. D. M. (1968). Slips of the tongue. *International Journal of Language and Communication Disorders, 3*, 2-12.

Brown-Schmidt, S., & Conseco-Gonzalez, E. (2004). Who do you love, your mother or your horse? An event-related brain potential analysis of tone processing in Mandarin Chinese. *Journal of Psycholinguistic Research, 33*, 103-135.

Chang, H.-C., Lee, H.-J., Tzeng, O. J. L., & Kuo, W.-J. (2014). Implicit target substitution and sequencing of lexical tone production in Chinese: An fRI study. *PLoS ONE, 9*, e83126.

Chao, Y. R. (1930). A system of tone letters. *Le Maître Phonétique, 45*, 24-27.

Chao, Y. R. (1947). *Cantonese primer*. Cambridge, MA: Harvard University Press.
Chen, J.-Y. (1999). The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue. *Applied Psycholinguistics, 20*, 289-301.
Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43*, 15-26.
Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by an implicit priming task. *Journal of Memory and Language, 46*, 751-781.
Chen, J.-Y., & Dell, G. S. (2006). Word-form encoding in Chinese speech production. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics (Vol. 1: Chinese)* (pp. 165-174). Cambridge, UK: Cambridge University Press.
Chen, J.-Y., O'Seaghdha, P. G., & Chen, T.-M. (2016). The primacy of abstract syllables in Chinese word production. *Journal of Experimental Psychology: Learning Memory and Cognition, 42*, 825–836. doi:http://doi.org/10.1037/a0039911
Chen, M. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge: Cambridge University Press.
Chen, T.-M., & Chen, J.-Y. (2013). The syllable as the proximate unit in Mandarin Chinese word production: An intrinsic or accidental property of the production system? *Psychonomic Bulletin & Review, 20*, 154-162.
Chen, T.-M., Dell, G. S., & Chen, J.-Y. (2004). A cross-linguistic study of phonological units: Syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English. *Cognitive Science Society, 26*, 216-220.
Cheung, K.-H. (1986). *The phonology of present-day Cantonese.* (Doctoral dissertation), University College London, London.
Costa, A., Alario, R.-X., & Sebastián-Gallés, N. (2007). Cross-linguistic research on language production. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 531-546). Oxford: Oxford University Press.
Cutler, A. (1980). Errors of stress and intonation. In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of tongue, ear, pen, and hand* (pp. 67-80). New York: Academic Press.
Dell, G. S. (1984). Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 222-233.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.
Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language, 27*, 124-142.
Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149-195.
Dell, G. S., Schwartz, M., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*, 801-838.
Duanmu, S. (1995). Metrical and tonal phonology of compounds in two Chinese dialects. *Language: Journal of the Linguistic Society of America, 71*(2), 225-259.
Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford: Oxford University Press.
Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry, 8*, 505-520.

Ferber, R. (1995). Reliability and validity of slip-of-the-tongue corpora: A methodological note. *Linguistics, 33*, 1169-1190.
Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30*, 139-162.
Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.
Fromkin, V. (Ed.) (1980). *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*. San Diego: Academic Press.
Gandour, J. (1977). Counterfeit tones in the speech of southern Thai bidialectals. *Lingua, 41*, 125-143.
Gandour, J. (1998). Aphasia in tone languages. In P. Coppens, Y. Lebrun, & A. Basso (Eds.), *Aphasia in atypical populations* (pp. 117-142). London: Lawrence Erlbaum.
Gandour, J., Xu, Y., Wong, D., Dzemidzic, M., Lowe, M., Li, X., & Tong, X. (2003). Neural correlates of segmental and tonal information in speech perception. *Human Brain Mapping, 20*, 185-200.
Gao, M. (2009). *Mandarin tones: An articulatory phonology account.* (Doctoral dissertation), Yale University, New Haven.
Garrett, M. (1984). The organization of processing structure for language production. In D. Caplan, A. R. Lecours, & A. Smith (Eds.), *Biological perspectives on language* (pp. 172-193). Cambridge, MA: MIT Press.
Glerean, E. (2014). Mantel test - Matlab implementation. Retrieved from https://doi.org/10.6084/m9.figshare.1008724.v3
Goldrick, M., & Blumstein, S. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649-683.
Goldrick, M., Folk, J., & Rapp, B. (2010). Mrs Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language, 62*, 113-134.
Griffin, Z. M., & Crew, C. M. (2012). Research in language production. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 409-425). Cambridge: Cambridge University Press.
Kember, H., Croot, K., & Patrick, E. (2015). Phonological encoding in Mandarin Chinese: Evidence from tongue twisters. *Language and Speech, 58*, 417-440.
Kuo, G. (2010). Production and perception of Taiwan Mandarin syllable contraction. *UCLA Working Papers in Phonetics, 108*, 1-34.
Lee, J. L. (2012). The representation of contour tones in Cantonese. *Berkeley Linguistics Society, 38*, 272-286.
Lee, K. G. (2003). *Syllable fusion in Cantonese connected speech.* (MPhi thesis), University of Hong Kong, Hong Kong.
Leung, M. T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computer, 36*, 500-505.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1-75.
Liu, J. H. C., & Wang, H. S. (2008). Speech errors of tone in Taiwanese. *North American Conference on Chinese Linguistics 20*, 189-203.
MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia, 8*, 323-350.

Matthews, S., & Yip, V. (2011). *Cantonese: A comprehensive grammar*. London: Routledge.
Meyer, A., S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29*, 524-545.
Mok, P. P.-K., Zuo, D., & Wong, P. W.-Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*, 314-370.
Moser, D. (1991). *Slips of the tongue and pen in Chinese (Sino-Platonic Papers, No. 22)*. Philadelphia: University of Pennsylvania.
Nooteboom, S. G. (1969). The tongue slips into patterns. In A. J. van Essen & A. A. van Raad (Eds.), *Leyden studies in linguistics and phonetics* (pp. 114-132). The Hague: Mouton.
O'Seaghdha, P. G. (2015). Across the great divide: Proximate units at the lexical-phonological interface. *Japanese Psychological Research, 57*, 4-21.
O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115*, 282-302.
Packard, J. (1986). Tone production deficits in nonfluent aphasic Chinese speech. *Brain and Language, 29*, 212-223.
Pérez, E., Santiago, J., Palma, A., & O'Seaghdha, P. G. (2007). Perceptual bias in speech error data collection: Insights from Spanich speech errors. *Journal of Psycholinguistic Research, 36*, 207-235.
Poser, W. (1984). *The phonetics and phonology of tone and intonation in Japanese.* (Doctoral dissertation), MIT, Cambridge, MA.
Qu, Q., Damian, M. F., & Kazanina, N. (2012). Sound-sized segments are significant for Mandarin speakers. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 14265–14270.
Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460-499.
Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). *Psychological Review, 111*, 561-572.
Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From Germanic languages to Mandarin Chinese and Japanese. *Japanese Psychological Research, 57*, 22–37. doi:http://doi.org/10.1111/jpr.12050
Selkirk, E., & Shen, T. (1990). Prosodic domains in Shanghai Chinese. In S. Inkelas & D. Zec (Eds.), *The Phonology-Syntax connection* (pp. 313-337). Chicago: University of Chicago Press.
Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Copper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Erlbaum.
Shattuck-Hufnagel, S. (1987). The role of word onset consonants in speech production planning: New evidence from speech error patterns. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 17-51). Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41-55.

Shen, J. (1993). Slips of the tongue and the syllable structure of Mandarin Chinese. In S.-C. Yau (Ed.), *Essays on the Chinese language by contemporary Chinese scholars* (pp. 139-161). Paris: Centre de Recherche Linguistiques sur L'Asie Orientale. École des Hautes Études en Sciences Sociales.

Shih, C.-L. (1986). *The prosodic domain of tone sandhi in Chinese.* (Doctoral dissertation), University of California, San Diego, San Diego.

Simner, J., Hung, W.-Y., & Shillcock, R. (2011). Synaesthesia in a logographic language: The colouring of Chinese characters in Pinyin/Bopomo spellings. *Consciousness and Cognition, 20*, 1376-1392.

Stemberger, J. P. (1982/1985). *The lexicon in a model of language production*. New York: Garland.

Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review*. Bloomington: Indiana University Linguistics Club.

Stemberger, J. P. (1989). Speech errors in early child language production. *Journal of Memory and Language, 28*, 164-188.

Stemberger, J. P. (1993). Spontaneous and evoked slips of the tongue. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies. An international handbook* (pp. 53-65). Berlin: Walter de Gruyter.

Van Linker, D., & Fromkin, V. (1973). Hemispheric specialization for pitch and "tone": Evidence from Thai. *Journal of Phonetics, 1*, 101-109.

Wan, I.-P. (2006). Tone errors in normal and aphasic speech in Mandarin. *Taiwan Journal of Linguistics, 4*, 85-112.

Wan, I.-P., & Jaeger, J. J. (1998). Speech errors and the representation of tone in Mandarin Chinese. *Phonology, 15*, 417-461.

Wang, J., Wong, A. W.-K., Wang, S., & Chen, H.-C. (2017). Primary phonological planning units in spoken word production are language-specific: Evidence from an ERP study. *Scientific Reports, 7*, 5815.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 387-398.

Wells, R. (1951). Predicting slips of the tongue. *Yale Scientific Magazine, 3*, 9-30.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76*, 1-15.

Wilshire, C. E. (1998). Serial order in phonological encoding: An exploration of the "word onset effect" using laboratory-induced errors. *Cognition, 68*, 143-166.

Wong, A. W.-K., & Chen, H.-C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1172-1190.

Wong, A. W.-K., & Chen, H.-C. (2009). What are effective phonological units in Cantonese spoken word planning? *Psychonomic Bulletin and Review, 16*, 888–892. doi:http://doi.org/10.3758/PBR.16.5.888

Wong, A. W.-K., & Chen, H.-C. (2015). Processing segmental and prosodic information in spoken word planning: Further evidence from Cantonese Chinese. *Japanese Psychological Research, 57*, 69-80. doi:http://doi.org/10.1111/jpr.12054

Wong, A. W.-K., Chiu, H.-C., Wang, J., Wong, S.-S., & Chen, H.-C. (2018). Electrophysiological evidence for the time course of syllabic and sub-syllabic encoding in Cantonese spoken word production. *Language, Cognition and Neuroscience*. doi:10.1080/23273798.2018.1562559

Wong, A. W.-K., Huang, J., & Chen, H.-C. (2012). Phonological units in spoken word production: Insights from Cantonese. *PLoS ONE, 7*, 1-10.

Wong, W. Y. P. (2006). *Syllable fusion in Hong Kong Cantonese connected speech.* (Doctoral dissertation), The Ohio State University, Columbus, OH.

Yip, M. (1992). Prosodic morphology in four Chinese dialects. *Journal of East Asian Linguistics, 1*, 1-35.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yue-Hashimoto, O.-k. (1972). *Phonology of Cantonese* (Vol. 1). Cambridge: University Press.

Zhang, Q., & Damian, M. F. (2019). Syllables constitute proximate units for Mandarin speakers: Electrophysiological evidence from a masked priming task. *Psychophysiology, 56*, e13317.

Zhang, Q., & Zhu, X. (2011). The temporal and spatial features of segmental and suprasegmental encoding during implicit picture naming: An event-related potential study. *Neuropsychologia, 49*, 3813–3825. doi:http://doi.org/10.1016/j.neuropsychologia.2011.09.040