UNIVERSITY OF CALIFORNIA

Los Angeles

Tone Sequences in Lexical Processing of Beijing Mandarin

A thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Linguistics

by

Isabelle Lin

2017

ABSTRACT OF THE THESIS

Tone Sequences in Lexical Processing of Beijing Mandarin

by

Isabelle Lin

Master of Arts in Linguistics

University of California, Los Angeles, 2017

Professor Kie Ross Zuraw, Co-Chair

Professor Megha Sundara, Co-Chair

When processing a string of syllables in Mandarin Chinese, tone information needs to be taken into account in order to identify possible words. However, most of the previous studies on processing of tone use monosyllables, which might underestimate the role of tone information. The purpose of this thesis is to examine the role of tone in contexts where it is maximally informative. To this end, we conduct a set of 3 experiments. Using corpus data, we show in experiment 1 that tone is more informative on disyllables than monosyllables. In experiment 2, we use existing corpus data as well as newly acquired speech data to define two alternative measures of tone frequency for disyllables: ditone sequence frequency (the likelihood of encountering a given sequence of two tones on a disyllabic word) and tone bigram frequency (the likelihood of encountering a given sequence of two tones in running speech, regardless of word boundaries). Based on these results, we investigate in

experiment 3 the role of tone sequences in disambiguating two segmentally identical disyllabic word candidates that differ only in tone. Using a priming paradigm, we presented native speakers of Mandarin with a disyllabic sequence that was tonally ambiguous between two lexical entries. We show that tone frequency plays a separate role from word frequency and interferes with word frequency information during the processing of disyllables in Mandarin. When word frequency and tone frequency did not favour the same candidate, a tonally-matched prime reduced the likelihood of picking the candidate with the matching tone. Results suggest that listeners are sensitive to the overall likelihood of encountering a given sequence of tones in running speech, regardless of word boundaries (tone bigram frequency).

The thesis of Isabelle Lin is approved.

Patricia Keating

Kie Ross Zuraw, Committee Co-Chair

Megha Sundara, Committee Co-Chair

University of California, Los Angeles

2017

Table of Contents

Table of Figures

List of Tables

Acknowledgements

This work was funded thanks to Student Research Support Committee of the UCLA Department of Linguistics. It was made possible thanks to the help and insights of many people, to whom I wish to express my most heartfelt thanks.

First and foremost, I am extremely grateful to my committee members, Megha Sundara, Kie Zuraw and Patricia Keating for their advice, comments, and much appreciated moral support during the entire project. I would also like to thank all the members of the Phonetics and Phonology seminars for their feedback.

I would like to thank my undergraduate research assistants, Dong Dai, Jin Weifeng, Wang Junyi, Xiong Yihan and Zeng Haixin for their patience in helping with the segmentation and annotation of the corpus recordings.

The members of the UCLA Statistical Consulting Center helped with data wrangling and subduing models that refused to converge.

Many thanks are due to my fellow graduate students in the department, who provided such an amazing environment in which to learn how to do research.

Lastly, I thank Olivier Wang for for his insights on programming, but most of all for his constant personal support.

All mistakes are my own.

Introduction

In Mandarin Chinese, any given syllable is defined by the combination of its segmental content and lexical tone. The tones are traditionally described as: high level tone (T1), rising tone (T2), low dipping tone (T3) and falling tone (T4), (Chao, 1948). In addition, some function words or syllables in weak prosodic position might bear a reduced tone (Chen, Xu, 2006), which we will refer to as T5. To recognize a word in Mandarin, it is presumably necessary to extract both segmental and suprasegmental information from the speech signal in order to retrieve the correct lexical entry. Previous studies on the role of tone in Mandarin suggest that tone information plays a minor role in speech processing. However, these results come mainly from monosyllabic words, in which tone might not be very informative. The purpose of this thesis is to find contexts in which tone is maximally informative, and examine in these specific contexts the contribution of tone to identifying words. We first summarize some of the previous results on processing of tone.

1. Tone information as secondary to segmental information

Previous studies show that native speakers of tone languages rely unequally on segmental and tone information. In a homophone decision task, where participants read Chinese characters either aloud or silently, Taft and Chen (1992) showed that Cantonese and Mandarin speakers were slower to decide whether words mismatching in tone only were homophonous (e.g. qu3 vs. qu4) than to make the same decision for words mismatching in vowel only (qi4 vs. qu4) or in both segmental and tonal content (e.g. nian2 vs. qu4). This suggests that segmental mismatches were detected faster than tone mismatches and used separately to decide on homophony, and perhaps that syllables mismatching in tone only are more likely to be mistaken as homophones. This preeminence of segmental information is also supported by Cutler and Chen (1997). Using an auditory lexical decision task

with Cantonese speakers, they found that non-words mismatching in one tone (e.g. /bok8-si2/) were more often mistakenly accepted as words if a real word existed with identical segmental content (e.g. /bok8-si6/) than if a real word existed with the same tone but a mismatch in one vowel (e.g. /bok8-sy6/). This again suggests that segmental content alone provides reliable information so that speakers can make a confident guess at the corresponding lexical entry.

However, Mandarin listeners do not ignore tone information altogether. Fox and Unkefer (1985) found a Ganong effect (Ganong, 1980) on Mandarin tone categorization: when presented with a synthetic monosyllabic token ambiguous between two tones, participants gave tone responses that would make it a real word. This suggests that lexical entries are stored with tone information and participants use this tone information from the lexicon to supplement insufficient tone information in the signal.

These results show that segmental information and pitch information coming from tone are separable, and that on monosyllables tone mismatches matter less than segmental mismatches. This preference for segmental information can also be found in the relative timelines of processing segments and tones.

2. Priming effects of tone

Not only is tone less informative than segments, it is also available later in the processing of auditory input. Tone information in Mandarin is mainly carried by F_0 (Howie, 1976, Gandour, 1983, Repp and Lin, 1990, Liu and Samuel, 2004). Because tone identification in Mandarin depends on a pitch contour, tonal information is available relatively late, towards the middle or late portions of the syllable (Howie, 1976, Whalen and Xu, 1992). This delay seems to be reflected in the timeline of processing tonal information. For instance, under a COHORT model of lexical retrieval (Marslen-

Wilson, 1984) in which an auditory input triggers first the activation of all possible word candidates (lexical access), followed by competition among said candidates before a unique item is retrieved (lexical selection), a difference in processing timelines might cause tonal information to be active at a later stage during lexical retrieval.

Consistent with this idea, in studies examining the perceptual processing of monosyllabic words, reaction times in lexical decision tasks suggest that while segmental information is processed early and plays the major role in lexical access, tonal information is processed later and is decisive only in lexical selection (Lee, 2007, Hu et al., 2012, Shuai et al., 2012). Event related potentials in detection of segment-induced vs. tone-induced semantic anomalies also show that segmental information is processed earlier than tone information (Brown-Schmidt and Canseco-Gonzalez, 2004).

In addition to being available at different timepoints, segmental and tonal information also differ in the type of priming effects they can elicit. Studies using priming paradigms have found facilitatory effects for matching segmental content (Lee, 2007, Sereno and Li, 2015). However, there is no clear consensus as to whether tone also induces such effects, and if it does, whether that effect is facilitatory or inhibitory. In a series of priming tasks, Lee (2007) found different facilitation effects depending on the ISI between prime and target. At 250ms, facilitation occurred only when prime and target were identical (e.g. lou2-lou2), and not with a segmental-only (lou2-lou3) or tone-only match (lou2-mang2). However, at 50ms, facilitation was found for the identical pair and also for the tone minimal pair (lou2-lou3), but still not for the tone-only match (lou2-mang2). Based on these results, he proposes that either tonal or segmental information can be active in blocking inappropriate lexical candidates at different times: segmental content is processed first in lexical activation, and all tone combinations are activated, then tone information reduces the number of candidates, still at a relatively early stage. By 250ms, candidates mismatching in tone have been

eliminated and priming no longer occurs between minimal tone pairs. Sereno and Li (2015)'s findings also support this processing timeline, though they found slightly different priming effects than Lee (2007). In an auditory lexical decision task with and ISI of 250 ms, they found a facilitatory priming effect between monosyllabic stimuli when only segmental information overlapped (ru3-ru4), but not when only tones matched (sha4-ru4).

However, also with an ISI of 250 ms, Poss, Hung and Will (2008) found that for monosyllabic Mandarin words and pronounceable nonwords, auditory primes mismatching in segmental content but matching in tone with the target (mo3-ba3) caused a delay in a lexical decision and a shadowing task. This result suggests that previous identification of a tone might inhibit the selection of a word candidate bearing the same tone. The authors propose that tone information might also induce the lexical activation of a group of words bearing the same tone, increasing competition with the target word and thus delaying lexical access.

Curiously, almost all these studies use Mandarin monosyllabic stimuli (except for Cutler and Chen (1997), for which stimuli were disyllabic, but in which only one syllable was compared). While it is true that most legal monosyllabic segments+tone combinations can form independent words/morphemes in Mandarin, every such combination can correspond to a number of homophones (e.g. mei2 can correspond to the morphemes 'charcoal', 'eyebrow', 'berry', 'mold', among many more possibilities: in modern Mandarin, these morphemes are not usually the words used in spoken language to refer to these meanings). A great number of the content words used in spoken language are instead disyllabic compounds, such that the other syllable of the compound disambiguates the target morpheme (e.g. mei2tan4 'charcoal', mei2mao2 'eyebrow', cao3mei2 'strawberry', fa1mei2 'to grow moldy'). Therefore, when processing a monosyllabic segments+tone combination like mei2 uttered to mean 'charcoal', complete identification of segmental and tone

information still does not bring the listener to identify a unique lexical entry, and the listener is left at a stage where at least 4 homophones are possible. Thus, studies using monosyllabic stimuli are likely to underestimate the contribution of tone information to Mandarin lexical access.

We examine corpus data in order to assess whether this is the case. In the first experiment, we use an existing corpus of Mandarin subtitles to show that tone information is more useful in uniquely identifying Mandarin disyllables rather than monosyllables. Based on this result, we determined to investigate the role of tone in disyllables. If present, such a role is likely to be mediated by frequency effects, but frequency for tone on disyllables can be measured in different ways. In the second experiment, we define two alternative measures of tone sequence frequency using data from an existing corpus as well as newly acquired semi-spontaneous speech data. In the third experiment, we examined the contribution of tone sequence frequency in disambiguating segmentally identical disyllabic words. Since segmental content is identical on the two competing lexical entries, this would allow us to focus more clearly on the influence of tone in biasing lexical selection. Because previous studies used monosyllabic stimuli and lexical decision, it is not clear that lexical selection of a unique word is achieved during the task. This is why we use a word identification task, in which participants must give an exact word response. We use a priming paradigm with tone sequences to show that tone sequence frequency plays a role independent from word frequency during lexical selection. In a smaller, pilot dataset, we also examined which of the alternative measures of tone sequence frequency defined in the second experiment is more appropriate to describe the role of tone in processing disyllabic words.

Experiment 1: Informativeness of tones in Mandarin

In Experiment 1, we compare the relative informativeness of tone for the recognition of Mandarin monosyllables vs. disyllables. One possible measure of informativeness is functional load. The functional load of a given contrast in a language L estimates the increased homophony induced by its neutralization. The higher the functional load of a contrast, the more information it contributes. It is likely that contrasts with a high functional load are more active in lexical activation and selection, and very probably processed in priority. Using an information-theoretic framework, Surendran and Niyogi (2003) defined a measure of the functional load of phonological contrasts. In their implementation, a language L is represented as a sequence of discrete units x, and Shannon's entropy H(L) is computed as:

$$H(L) = -\sum_x p(x) log_2 p(x) \quad (1)$$

Entropy is then computed for a hypothetical language L' in which a given contrast is inactive. For instance, Mandarin without tone contrasts would be a language in which tone has been removed, so that syllables differing only in tone (e.g. ma1 and ma2) are indistinguishable, but syllables differing by segmental content are distinct (e.g. ma and mi, with any tone). The functional load of a given contrast is thus defined as the relative difference (in percentage) in entropy between L and L'[1]:

$$FL = \frac{H(L)-H(L')}{H(L)} \quad (2)$$

[1] Under this definition, functional load quantifies information loss. The greater the functional load of a contrast, the more its loss impacts the language. Therefore, the functional load of the set of all contrasts in the language (L=L') is null, since no information is lost.

Surendran and Levow (2004) computed a measure of the functional load of tone, consonants and vowels in Mandarin, and found tone to be as informative as vowels in Mandarin, and both tone and vowel to be less informative than consonants. In a subsequent cross-linguistic comparison, Oh and colleagues (2013) examined those same parameters for Cantonese, English, Japanese, Korean and Mandarin, and confirmed this result (Figure 1 below).

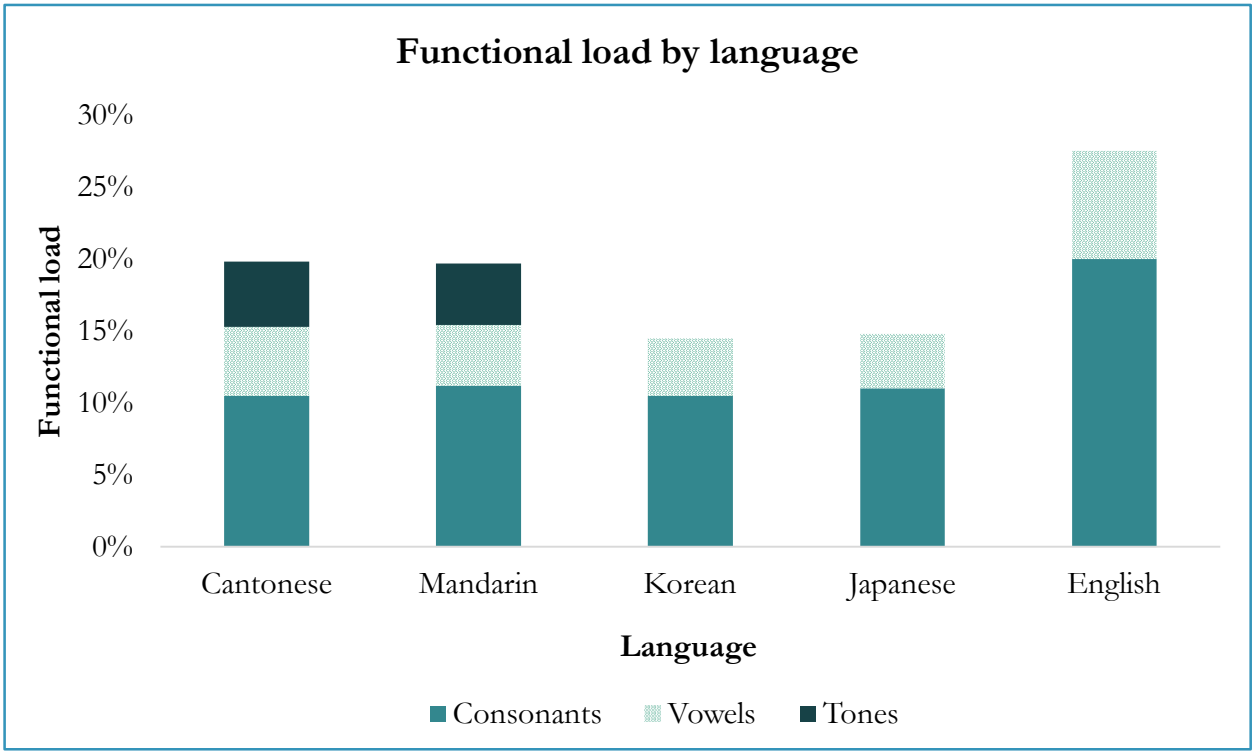

Figure 1. From Oh, Pellegrino, Coupé and Marsico (2013). Functional load carried by each phonological subsystem (Vocalic V, Consonantal C, and Tonal T).

These results confirm the experimental findings that segmental content (or consonants and vowels taken together), is more informative than tones in Mandarin. Vowels on their own are about as informative as tones in Mandarin, but we will not consider their contribution separately here and rather take the functional load of Vowels + Consonants to represent the functional load of segments. However, because entropy is computed using the syllable as a comparison unit in which segments and tone are specified, and the language L is generated using a random (stationary ergodic) process,

these calculations will not allow us to compare the role of segments and tone in words of different lengths.

Using the same basic principle of loss of distinctiveness, we compare the number of monosyllabic and disyllabic homophones. To confirm whether there really are more monosyllabic homophones than disyllabic ones in Mandarin, we examined the number of homophones in the lexicon. In homophones, the combination of segmental and tonal information is insufficient to select a unique lexical entry. If there are more monosyllabic homophones, then studies using monosyllabic stimuli will underestimate the contribution of tone information to identifying words since lexical selection cannot always be completed with isolated monosyllables.

1. Methods

We used the SUBTLEX-CH corpus (Cai and Brysbaert, 2010) as the basis for this analysis. This is the largest corpus of Standard Mandarin that is freely available, containing data for 46.8 million characters and 33.5 million words. Based on film and television subtitle databases, the main purpose of this corpus is to provide word and character frequencies, different from our purpose here, which is simply to compile a modern lexicon. However, we chose to use the SUBTLEX-CH rather than dictionaries, because the latter heavily overestimate the proportion of monosyllabic words as they include a vast amount of archaic monosyllabic words that are no longer used in modern spoken speech. We use type frequencies from the SUBTLEX-CH corpus as an approximation of a realistic modern Mandarin lexicon. In the following analysis, we take percentages of the corpus to be a proxy for percentages of the lexicon.

Pinyin (romanized spelling of Mandarin segmental content) and tone information were not directly provided in the original corpus. Using a Python script, we tagged the corpus for pinyin and tones

using the CC-CEDICT dictionary database. For each word in the SUBTLEX-CH, the script looked up the entry matching in orthography and part of speech in the dictionary, and added tags for pinyin and tone to the corpus. For monosyllables, a given character can sometimes have different readings, but these readings often correspond to different parts of speech, so matching part of speech can solve this issue in part. If the different readings did not correspond to different parts of speech, the script checked if the corpus contained another instance of the same character and same part of speech. If this was the case, the first reading from the dictionary (presumably the more frequent reading) was assigned to the version of the character that was tagged with the higher word frequency, and the subsequent readings assigned to the other instances of the character in order of decreasing frequency. In case a word in the corpus could not be found in the dictionary (e.g. proper names), the script would search for the individual characters in the word and tagged the corpus with the reading obtained from concatenating the readings from individual characters.

Using the corpus tagged for pinyin and tone, we computed the number of homophones corresponding to (1) the same pinyin (segmental content) and (2) the same pinyin+tone combination (segmental + tone content) in both monosyllabic and disyllabic words.

Results and discussion

Monosyllabic words constitute 11.15% of the corpus (of types). Figure 2 on the following page shows the percentage of the corpus for which a combination of segments (dark bars) or segments + tone (light bars) selects a certain number of homophones. For instance, combinations of segments that select a unique word represent 0.64% of the corpus (leftmost dark bar). This represents 5.7% of monosyllables. For a given combination of segments, there are on average 12 homophones (average of dark bars).

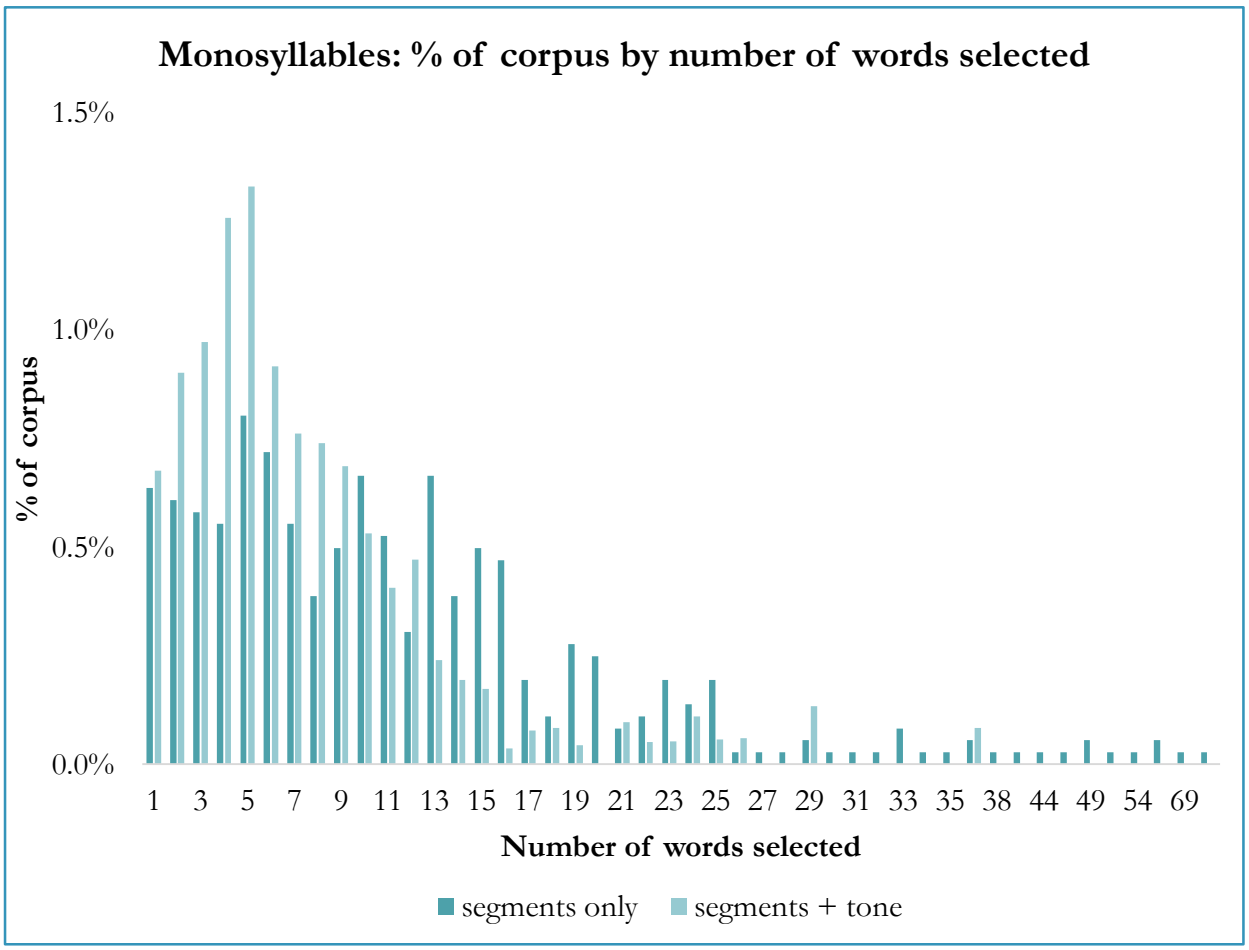

Figure 2. Percentage of corpus by number of monosyllabic words selected using segments only or both segments and tones. Monosyllables constitute 11.15% of the corpus. Combinations of segments that select a unique word represent 0.64% of the corpus (5.7% of monosyllables). Combinations of segments + tone that select a unique word represent 0.68% of the corpus (5.9% of monosyllables).

If we add in tone information (light bars), we find that each combination of segments + tone selects on average 4 homophones (average of light bars), which is an improvement from segments-only sequences. We note also that the maximal number of homophonous lexical items has decreased from 75 (rightmost dark bar) to 36 (rightmost light bar). However, the combinations that define a unique lexical item still constitute only 0.68% of the corpus (leftmost light bar), which is still very limited. From this, we can conclude that although tone information is helpful in narrowing down the number of candidates for monosyllabic words, its contribution is not sufficient to reach a uniquely defined lexical item in most situations.

Disyllabic words constitute 65.56 % of the corpus. Figure 3 below shows the percentage of the corpus for which a combination of segments (dark bars) or segments + tone (light bars) selects a certain number of homophones. If we take the corpus to be an approximation of the lexicon and accept its word segmentations, this confirms that Mandarin has more disyllabic than monosyllabic words. In disyllabic words, the same sequence of segments selects on average a single lexical item (average of dark bars = 1.53). The pinyin sequences that define a unique lexical item constitute 51.26% of the corpus (leftmost dark bar). This means that segmental content alone suffices to pinpoint a single lexical item for slightly over half of the lexicon.

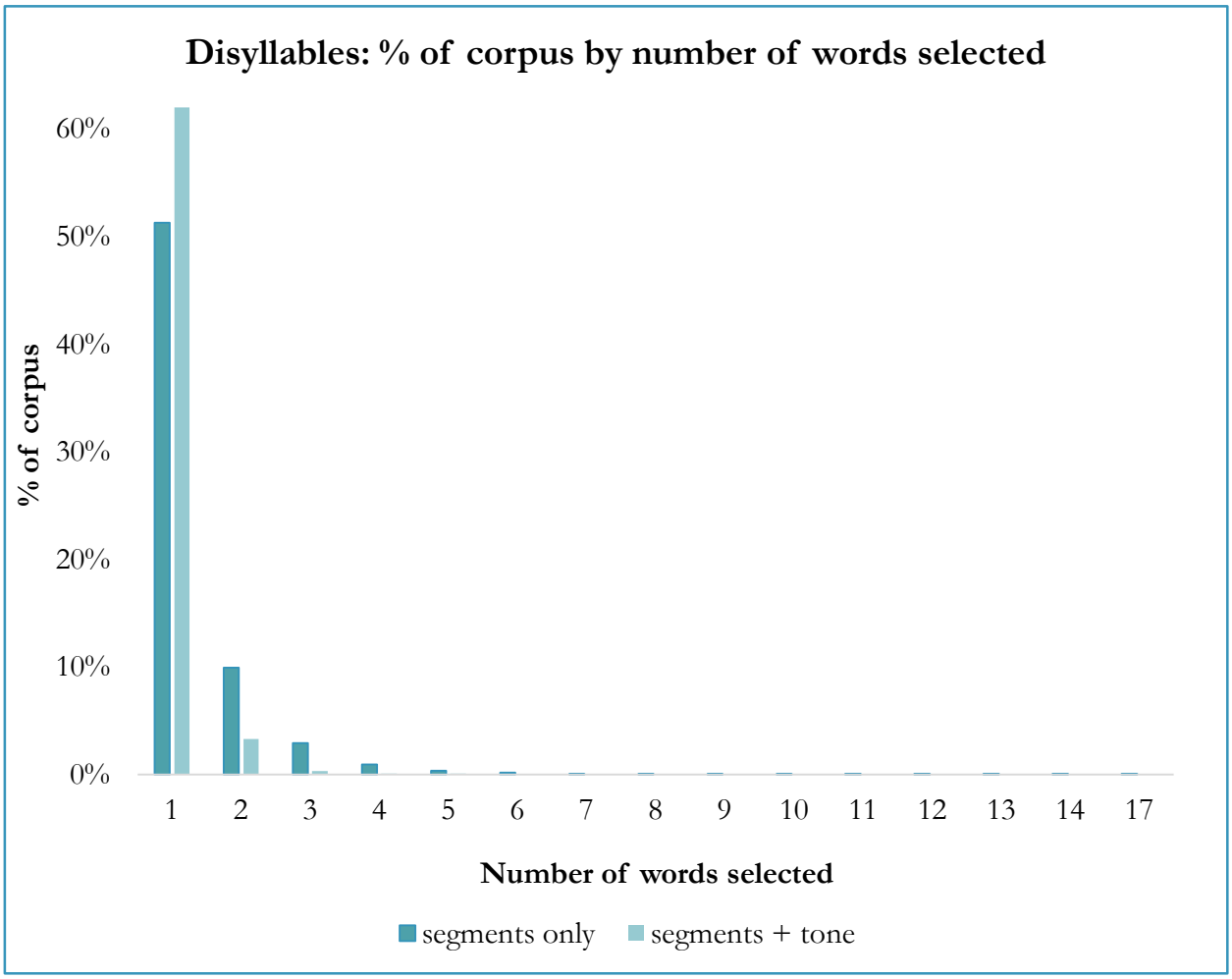

Figure 3. Percentage of corpus by number of disyllabic words selected using segments only or both segments and tones. Disyllables constitute 65.56 % of the corpus. Combinations of segments that select a unique word represent 51.26% of the corpus (78% of disyllables). Combinations of segments + tone that select a unique word represent 61.94% of the corpus (94% of disyllables)

When adding in tone information on disyllabic words, we find that each combination of segments + tone sequences has on average 1 homophone (average of light bars = 1.23), which is not a clear improvement compared to the contribution of segmental content only. We note however that the maximal number of homophonous lexical items has decreased from 17 (rightmost dark bar) to 5 (rightmost light bar). More importantly, the combinations of segments + tone that define a unique lexical item now constitute 62% of the corpus (leftmost light bar), which does show some improvement on the contribution of segmental information alone (compare to leftmost dark bar). For those sequences that still have homophones after the addition of tone, we note that they mostly have either 2 homophones (10.94% of the corpus, second light bar) or 3 (3.28% of the corpus, third light bar).

To determine whether adding tone information to segmental information reduces the number of homophones to the same extent for monosyllables and disyllables, we compared the contribution of tonal information in monosyllabic and disyllabic words using a Fisher's exact test in Stata 13 (StataCorp., 2013). This revealed that the number of words with homophones after segments + tone have been specified significantly differed by word length ($p < 0.001$, see Table 1 for contingency table). With pinyin + tone specified on monosyllables, more homophones were observed than expected under the null hypothesis. Conversely, less homophones than expected were observed for the disyllables. This suggests that tone contributes more in defining unique disyllables than monosyllables.

Table 1. Contingency table for Fisher's exact test to compare the contribution of tone to the selecting of unique lexical items for monosyllabic and disyllabic words. For monosyllables, 292 words can be defined using segments + tone and 4529 still have homophones. Homophones sets such as 'mei2' in the morphemes 'charcoal', 'eyebrow', 'berry', 'mold', are counted only once as 'mei2' having homophones. Values expected under the null hypothesis are presented in parentheses. The number of words with homophones after segments + tone have been specified significantly differed by word length ($p < 0.001$)

	Have no homophones	Have homophones	**Total**
monosyllables	292 (4,221.06)	4,529 (599.94)	4,821
disyllables	41,944 (38,014.94)	1,474 (5,403.06)	43,418
Total	42,236	6,003	48,239

Another aspect, that we can only compare qualitatively, is the maximal number of possible homophones for a given sequence of segments or segments + tone. Disyllabic sequences, with or without tone, have overall fewer homophones (see Table 2 below).

Table 2. Contribution of tone to the reducing of the maximal number of possible homophones for monosyllabic and disyllabic words. For monosyllables, the maximal number of homophones that a given sequence of segments can have is 75, and for a given sequence of segments + tone, 36.

	Maximal number of homophones pinyin only	Maximal number of homophones pinyin + tone	**Total**
monosyllables	75	36	111
disyllables	17	5	22

Given these results, looking at monosyllables would lead to underestimating the informativeness of tone in Mandarin. Therefore, we will focus on the case of disyllables to investigate the role of tone in lexical selection.

In particular, there are some disyllabic items where the informativeness of tone is crucial (10.94% of the corpus). This is the subset of segmental sequences where only tone can disambiguate between two segmentally identical homophones. If the preferred strategy for processing disyllabic words is to rely first on segmental content alone, then at the lexical selection stage all segmentally identical homophones are activated. Among these segmentally homophonous candidates, only tone information allows the completion of lexical selection, so this should be the best test case to investigate the role of tone. We do this in Experiment 3.

One possibility is that the effect of tone on Mandarin lexical access is modulated by how likely it is to encounter a certain sequence of tones in the language. Speakers are sensitive to various types of frequency patterns in the language (Ellis, 2002 review article). Overall, more frequent words are processed faster than less frequent ones. Given that we are interested in tone on disyllabic words, we need to examine frequency for sequences of two tones rather than individual tones. There are at least two possible ways in which one can measure frequency of tone sequences in disyllables, which we do in Experiment 2.

Experiment 2: Estimating frequency of tone sequences

When considering what would be an appropriate measure of tone sequence frequency in a language, a first possibility is that listeners have tone sequence representations that match disyllabic words, which we will refer henceforth as ditone sequences . Then the ditone sequence frequency is the likelihood of encountering a given sequence of two tones, T_aT_b, on a disyllabic word in Mandarin.

This can be calculated based on the SUBTLEX-CH (Cai and Brysbaert, 2010) that we used in the previous corpus study.

Another possibility is that listeners are sensitive to the general likelihood of encountering a sequence of two given tones in running speech, which we will refer to as tone bigram frequencies. This type of frequency cannot be calculated based on existing corpora: since the main purpose of the SUBTLEX-CH corpus is to provide word and character frequencies, it does not preserve the sentence structures surrounding each word. This is why we can use this data to extract ditone sequence frequencies, but not bigram frequencies in running speech (across word boundaries). Similarly, Google Ngrams (Google, Inc., Michel et al., 2011) does not contain this information either as it is based on written materials, which tends to be more careful than casual speech. Large corpora of spoken Mandarin, such as CALLHOME (McEnery and Xiao, 2008) do exist, but these resources are not freely accessible. Thus, we acquired new speech data in order to access tone bigram frequencies.

In order to obtain tone bigram frequencies, we collected semi-spontaneous speech data in Standard Mandarin. In the future, we aim to build an open-source annotated corpus of Mandarin for linguistic research. For the purposes of our tone sequence project, 5 hours (2 speakers per hour) of semi-spontaneous conversation were acquired and annotated.

We considered various possible methods of eliciting casual speech, starting with those used in sociolinguistic interviews. Labov (1972) suggested 3 methods of eliciting casual speech from speakers who are aware that their speech is being observed. One method involves getting participants to speak before and after specific speech-related tasks. Speakers will tend to assume that their speech is not examined outside of these tasks and thus produce more natural speech. We did not use this method as the amount of natural speech elicited by this method is very small.

Another method consists of prompting participants to speak about topics that would get them in a very emotional state. This relies on the assumption that speakers will be less aware of their speech when in a highly emotional state. Of course, the possible topics would depend on the cultural background and personal experiences of the speaker. This method would also allow for collection of larger amounts of data, since speakers would be more likely to speak at length about such topics. However, this method would also raise some issues for our project. Since these recordings would ultimately be made available in a corpus project, participants would likely be less willing to contribute emotional speech, which tends to be more personal. We could allow participants to choose which portions of the recordings they were willing to keep in the corpus, but this would lead to losing some unpredictable amount of speech time.

Labov's third suggestion is the one we eventually based our methods on. This involves getting participants to talk to close friends or colleagues, thus partially recreating a conversational setting that is familiar and comfortable to them. Our choice was also inspired by the methods used by Ernestus (2000) to build a corpus of casual Dutch. Ernestus recorded pairs of mutually acquainted speakers in a variety of speech tasks, including free conversation in the presence of the experimenter. Speakers were seated in the same room, and an experimenter acquainted with both participants would provide remarks or ask questions to animate the conversation. Based on the aims of the current project, we implemented a modified version of the free conversation task to elicit our corpus data. Our participants were not acquainted with the experimenter, and we expected that having the experimenter in the room would make them less likely to converse in a natural manner. They were therefore left on their own during the recording. Additionally, group recordings are difficult to transcribe and annotate due to speaker overlaps.

1. Methods

Participants

We recorded speech from 9 native speakers of Standard Mandarin from Beijing, China. 5 speakers were female, 4 male, and their ages ranged from 19 to 53. They were recruited through the recording studio staff, with prespecified criteria. Participants had to be adult native speakers of Standard Mandarin with no known history of hearing or speech impairment, able to read Chinese characters (elementary school completion level) and crucially had to bring with them a close friend who also fit these criteria. Five pairs of speakers were initially recruited, but one participant could not come to the recording, so one of the previous speakers took part in an additional conversation. We specified that participants should be from Beijing in order to limit effects of dialectal variations: though Standard Mandarin is spoken throughout mainland China in addition to the regional dialects, influences of other dialects are still detectable. We did not specify criteria for socioeconomic background, however all participants recruited were university students and office workers in small organizations (<100 people). Minimal reading ability was needed to use the conversation prompts. Since recruiting was finalized before recording started, we adapted conversation prompts to likely conversation topics given our participant population.

Recording

Because of the acoustic quality desired for the recordings and issues of administrative approval, we conducted recordings in a studio specialized in spoken word recording. Recordings were done at the JiuGongDaCheng Studio in Beijing, in 2 adjoining soundproof rooms. Participants sat at a desk facing a laptop, on which the conversation prompts were available as Powerpoint slides. The participant seated in one room could hear the other participant though headphones. Direct to disk recordings were made using a pair of CAD U37 USB Studio Condenser Recording Microphones,

placed on the desk where the participant sat. The audio recording and editing software Audacity was used as the recording program, with a sampling rate of 44100 Hz. Recording sessions were scheduled over 2 days. Each session lasted one hour.

Since we expected them to need some time at the beginning of the recording session to grow accustomed to the studio setting and start talking more naturally, we decided that 1 hour sessions would be adequate to let speakers adjust to the situation and produce a good amount of utterances. Since speakers were recorded on completely different tracks, 2 hours of recording were obtained for each session, 1 hour per speaker. As expected, a great amount of overlap occurred, and on average 37 minutes of usable utterances could be extracted for each speaker.

Participants were asked to engage in casual conversation, as if they were talking over the phone. This would mimic a typical situation in which they would be talking without seeing the other person. They were encouraged to pick their own favorite conversation topics, but warned that the recordings would be publicly available for researchers, though identifying data would not appear in the corpus. In case participants ran short of conversation topics, they were provided with a list of 45 possible topics presented in the form of PowerPoint slides on a laptop. They were told that those were only suggestions, and deviating to other topics was perfectly acceptable. These topics were given using as general terms as possible, to allow for more varied interpretations by the speakers. All speakers received the same topics, though in a different random order (same order for each pair of speakers). Topics included for example "traffic in Beijing", "recently watched TV show", "travelling", etc. (full list in Appendix I). They were chosen to be likely conversation topics given the population of speakers that was recruited.

2. Segmentation and annotation

Recordings of the semi-spontaneous dialogues were segmented manually by 6 native speakers of Standard Mandarin trained in speech segmentation. This was done using the ELAN Multimedia Annotator software (Max Planck Institute for Psycholinguistics). The software was chosen for its detailed transcription-oriented interface, as well as its compatibility with multiple data formats. The results of this manual segmentation will be used later to constitute the training data for automated audio segmentation using forced alignment. Annotation was manual for the transcription in simplified Chinese characters, indication of pinyin (alphabetic transcription convention for Mandarin) and tone (1-4, neutral tone was indicated as 5), and English translation at the sentence level. 20% of segmentation and annotation were cross-checked among transcribers. Five transcribers each checked a different part of the material, and one transcriber checked all cases of disagreement, which were set to the value that 2 out of 3 transcribers agreed on (93.2% agreement on position of boundaries and 98.4% agreement on content of transcription)

Annotation for part of speech and gloss were then automated using a Python script to search the CC-CEDICT Chinese to English dictionary (Denisowski, 1997). The script used the first translation matching pinyin and tone for a give character/compound. This approximation inevitably includes some errors, but as our current main focus is on segmental content and tone, we leave this issue to be solved at a later date.

3. Data architecture

ELAN can be used to build a data structure containing several types and hierarchical levels of information. In our segmentation and annotation system, the smallest unit of segmentation is

currently the syllable, corresponding to the single orthographic character. Annotation is done in UTF-8 encoding to allow simplified Chinese script. This level is used for 4 tiers:

- CNscript (ELAN linguistic type: practical orthography) contains the transcription in simplified Chinese characters. This is the main level of segmentation.
- Pinyin (ELAN linguistic type: detailed transcript) is a daughter tier to CNscript. All segments in CNscript are present, and annotated for segmental transcription in pinyin
- Tone (ELAN linguistic type: detailed transcript) is also a daughter tier to CNscript and contains transcription of surface lexical tone (1-4, neutral tone is coded as 5). Syllables that undergo tone 3 sandhi, a process by which the first of two syllables bearing tone 3 will be produced with a surface tone 2, is coded as a sequence of 2 and 3 rather than 3 and 3, and indicated in the sandhi tier.
- Sandhi (ELAN linguistic type: detailed transcript) is also a daughter tier to CNscript and indicates syllables that have undergone tone 3 sandhi (coded as S)

The next unit of segmentation is the word level. This level only includes some of the boundaries set in the syllable level, as multiple syllables can belong to a single word. Word units are usually between one and three syllables long, but can be longer on occasion (loans, proper names). The choice of boundaries to remove from the syllable level is done manually, but annotation is currently semi-automatic (matching items from CC-CEDICT dictionary, manual addition for missing entries). This level is used for 2 tiers:

- Gloss (ELAN linguistic type: detailed transcript) provides the translation in English (first translation matching pinyin and tone for a give character/compound)

- PoS (ELAN linguistic type: detailed transcript) is a daughter tier to Gloss and indicates the grammatical function of the corresponding unit (again, matching from the first pinyin and tone match for a give character/compound)

The last unit of segmentation is the sentence level. This level includes the fewest boundaries, marking only sentence-type units. This level is used for one tier only:

- ENGtrans (ELAN linguistic type: free translation) provides the translation in English at the sentence level. Annotations on this tier are manual.

4. Results and discussion

The current recordings allowed us to build a corpus of 22571 syllables, containing 4120 words, among which 518 were monosyllabic and 3471 were disyllabic words. The 22571 consecutive syllables constitute 22570 tone bigrams. This corpus is much smaller than the SUBTLEX-CH (46.8 million syllables), and because of the limited amount of speakers recorded and the use of conversation prompts, its lexical make-up might be skewed towards particular topics. If some words recur too often, the tone sequence distribution in the corpus might be skewed as a result. To check whether frequency measures calculated from this corpus might be biased because of such a skew, we compared the distribution of ditone sequence frequencies (within word frequencies) in the SUBTLEX-CH and in our corpus. We used a two-sample Kolmogorov-Smirnov test for equality of distribution functions in Stata 13. The two distributions were not significantly different ($p = 0.6$). Despite its smaller size, our corpus does not seem to overly distort the distribution of tone sequences in disyllabic words observed in the larger corpus. This means that if we find that ditone sequences and tone bigrams in our corpus provide significantly different definitions of tone sequence frequency, this difference should not arise from the lexical bias inherent in our data. Figure 4 below shows the comparison between the two distributions. We note that sequences containing

T5 are over-represented in our corpus: this is probably because words ending in T5 are more common in casual speech (diminutives, reduplication).

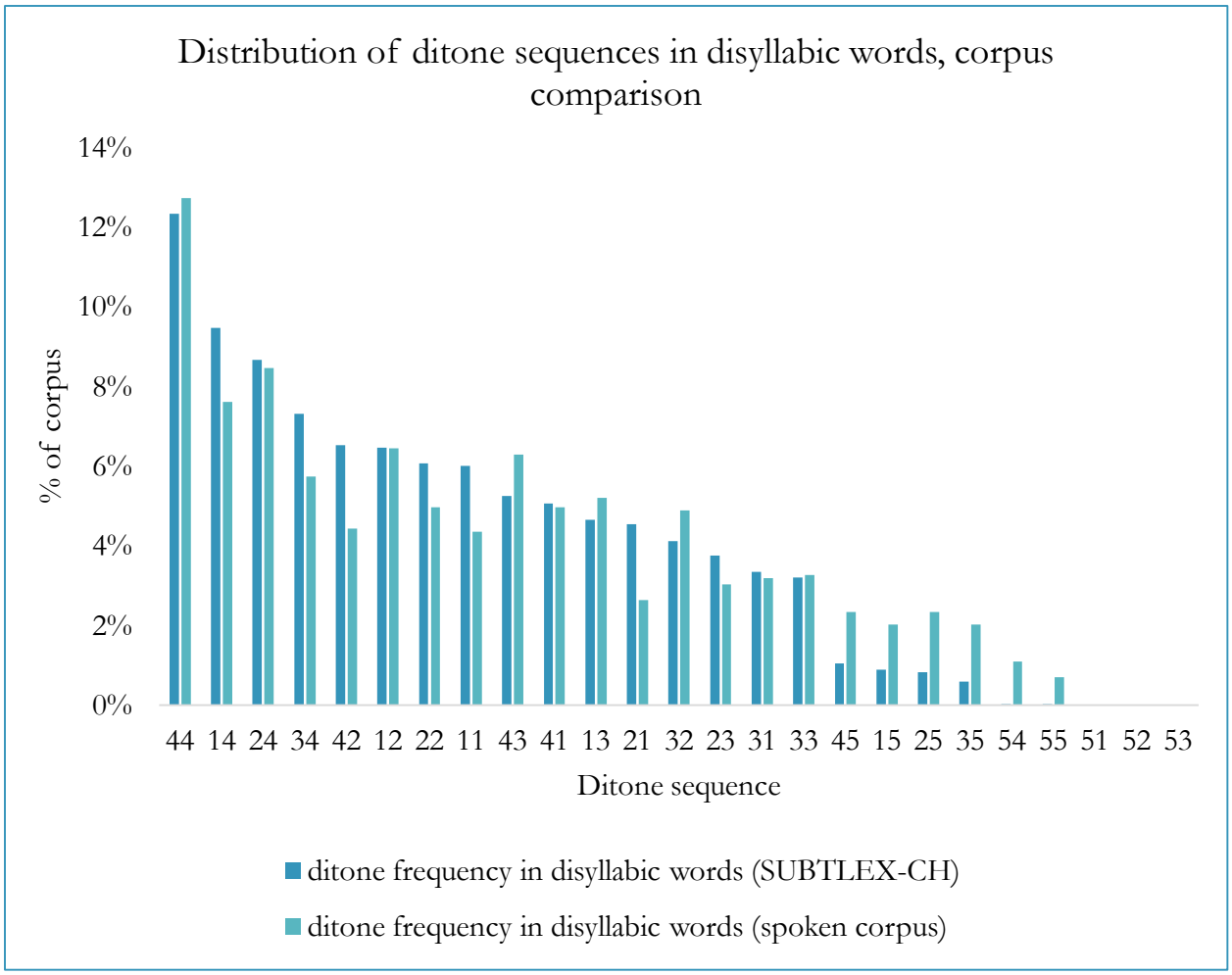

Figure 4. Distribution of ditone sequences in disyllabic words, corpus comparison. Distribution of all possible ditone sequences in the two corpora, in decreasing order from most frequent to least frequent in SUBTLEX-CH. We aimed to check whether the distribution of ditone sequences computed from our corpus of semi-spontaneous Mandarin (22571 syllables) showed skewness compared to the distribution obtained from SUBTLEX-CH. To allow for a comparison with the data obtained from the much larger SUBTLEX-CH (46.8 million syllables), we chose to use percentages rather than words per million. 44 means a sequence of T4T4.

We then computed the tone bigram frequencies (containing within word and across word sequences) from our corpus. This distribution of frequencies is shown in Figure 5 on the next page:

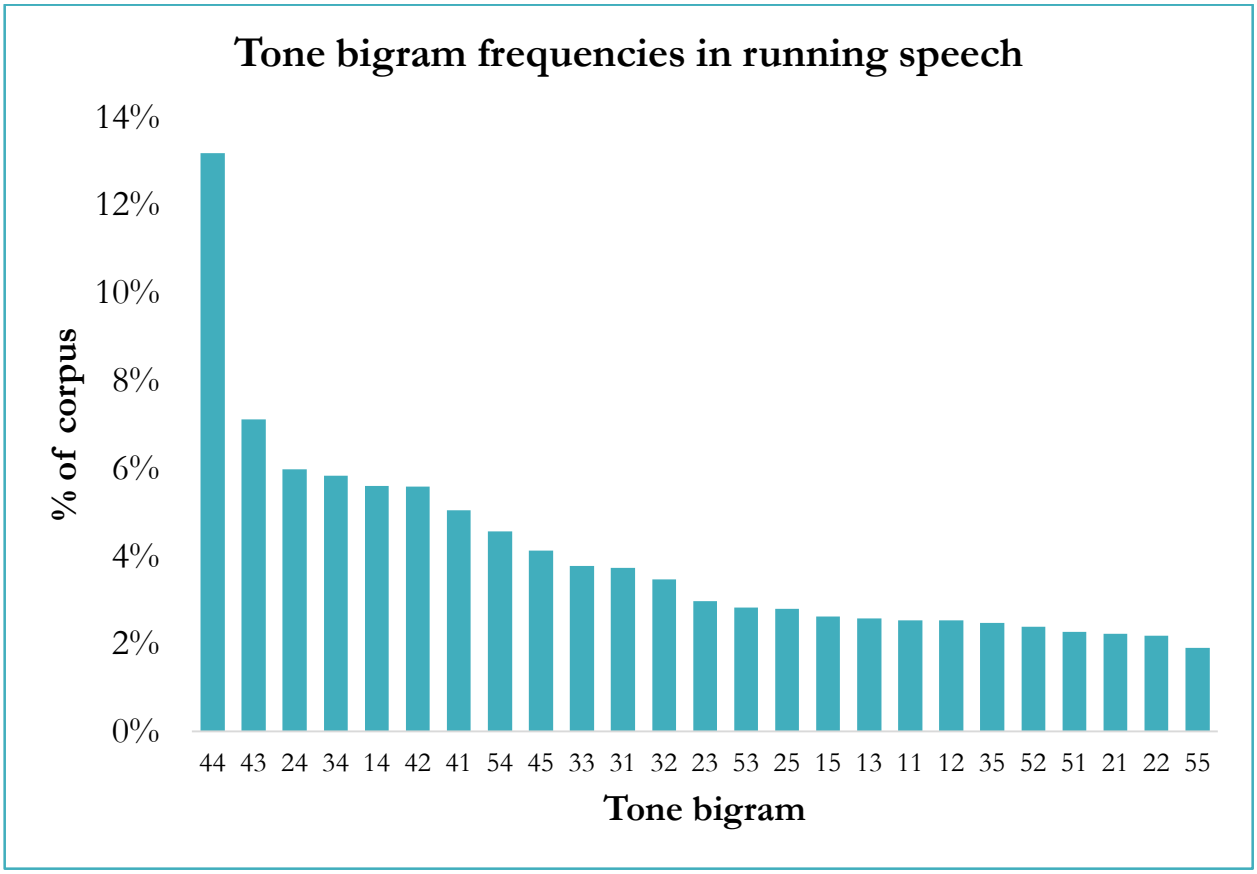

Figure 5. Tone bigram frequencies from corpus of semi-spontaneous Mandarin, computed from 22571 syllables. To allow for a comparison with the data obtained from the much larger SUBTLEX-CH (46.8 million syllables), we chose to use percentages rather than words per million. 44 means a sequence of T4T4.

As expected, the order of bigram frequencies obtained from our corpus differs from the ditone frequencies computed from the SUBTLEX-CH (within words). Setting apart the difference in size between the two databases (46.8 million syllables in the SUBTLEX-CH vs. 22571 syllables in our data), the frequencies should differ because natural speech does not consist of sequences of disyllabic words. In particular, some bigram sequences cannot appear in words/compounds: this is the case for most sequences starting with the reduced tone 5, which usually occur word/utterance finally[1]. Figure 6 on the next page shows the comparison of frequencies found using the

[1] SUBTLEX-CH contains exactly 2 cases of words bearing tones 54 or 55, both phrase-final functional expressions (的话 de5hua4, an optional hypothesis marker used in "if…then" constructions to mark the end

SUBTLEX-CH (ditone frequencies in words) and our corpus (tone bigram frequencies in running speech). A two-sample Kolmogorov-Smirnov test for equality of distribution functions show that the distributions as defined by ditone and bigram measures were significantly different ($p < .05$)[1]. Note that we are not necessarily examining to what degree these two distributions differ: this comparison is mainly to assess whether it could be reasonable to build alternative measures of tone frequency based on them.

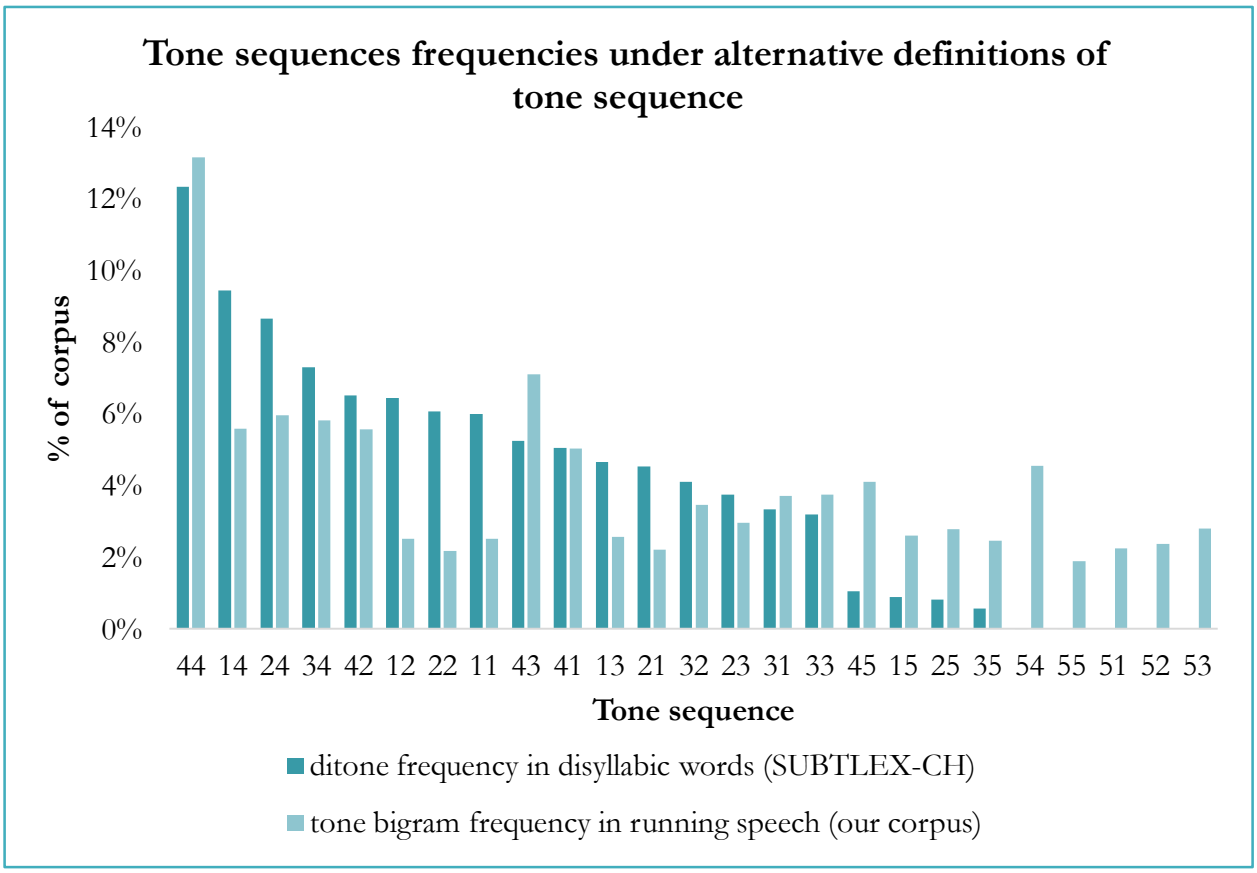

Figure 6. Tone sequence frequencies measured using alternative definitions of tone sequence. Darker bars: ditone frequencies in words obtained from the SUBTLEX-CH, ordered from most frequent to least frequent combination. Lighter bars: tone bigram frequencies obtained from our corpus, following the frequency order from SUBTLEX-CH to allow for better comparison.

---

of the hypothetical clause, and 着呢 zhe5ne5, a lexicalized combination of continuous aspect marker + focus marker)

[1] This is the *p*-value for the combined Kolmogorov-Smirnov test, corrected for small samples (n < 50) as we only have 25 frequencies to compare per type of measure.

These differences observed in the tone sequence distributions depending on the definition chosen (ditone vs. bigram) mean that they do indeed provide alternative measures of tone sequence frequency.

Even though there was no significant difference in the distribution of ditone frequencies calculated form the SUBTLEX-CH and from our new corpus, we chose to use the SUBTLEX-CH to compute ditone frequencies themselves, because the larger size of the corpus makes for a more accurate representation of the lexicon. Recall however that bigram frequencies cannot be extracted from the SUBTLEX-CH, and were calculated from our corpus.

In the next experiment, we tested Mandarin listeners' recognition of disyllabic minimal pairs when they are biased by tones matching one or the other candidate. Tone frequency was indexed using ditone vs. bigram frequencies to determine the appropriate measure.

Experiment 3: Word recognition experiment

In this experiment, we examined the role of tone sequences in lexical processing of Beijing Mandarin. We examined whether tone sequences can influence lexical selection in audio stimuli manipulated to be tonally ambiguous. Specifically, we tested whether a prior presentation of a word matching the tone pattern of one of the possible word candidates but otherwise unrelated to those candidates would influence a listener's identification of an ambiguous disyllabic item as one of the word candidates (e.g. T4T2 prime, matching ju4li2 'distance' or T3T4 prime, matching ju3li4 'to give an example'). Previous experiments focused on tone for monosyllables, but based on results from our corpus studies, tone is likely to play a lesser role in the recognition of Mandarin monosyllabic words. Thus, we investigated it in disyllabic words, where tone is more informative. Specifically, in the test

items tone disambiguates two segmentally identical homophones – item like this form about 10.94% of the corpus.

We used a word recognition task, in which participants must indicate exactly what word they recognized, in order to ensure that lexical selection has been completed. Note that in a lexical decision task, where participants only need to say whether they heard or saw a real word, the presence of homophones does not allow the identification of the unique lexical entry.

Overall, we expected participants to be more inclined to choose the more frequent word alternative. However, if tone sequences are decisive in lexical selection, we expected that the tone sequence on the prime should also influence the identification of the ambiguous target as one or the other word candidate. Possible frequency conditions are detailed in Table 3 on the next page. For items where the more frequent word bears the more frequent tone sequence (condition A), the more frequent word should be maximally salient during lexical access. In such cases, there is no way to disambiguate between the effects of tone or word frequency.

However, for conditions B, C and D the more frequent word did not bear the more frequent tone sequence. If word frequency favours one candidate while tone sequence frequency points to the other (condition D), participants' preference for the more frequent word should not be as strong, and therefore such items are most likely to be sensitive to tone priming.

Table 3. Possible frequency conditions. For condition B, the more frequent word (e.g. zaofan 34 'breakfast' is more frequent than 43 'to revolt') bears the more frequent tone pattern (frequency match) according to ditone frequency (34 more frequent than 43 on disyllabic words), but this pattern is the less frequent alternative (frequency mismatch) according to bigram frequency (in running speech, 43 is more frequent sequence than 34).

A	**More frequent ditone (match)** More frequent bigram (match)	***B***	Less frequent ditone (mismatch) More frequent bigram (match)
C	**More frequent ditone (match)** Less frequent bigram (mismatch)	***D***	Less frequent ditone (mismatch) Less frequent bigram (mismatch)

If tone sequence frequency does play a role during lexical selection, one can ask which of the two alternative measures of tone sequence frequency best accounts for that role. If frequency information on tone sequences is accessed during the task, it is possible that only a specific subset of that information is used, namely the sequence frequencies in disyllabic words. In this case, ditone frequency in words would be the better predictor of participant responses (condition B). Another possibility would be that listeners are more sensitive to the overall probability of encountering two given tones in a sequence than to disyllable-specific, lexicon-internal probabilities (condition C). In that case, tone bigram frequency in running speech would be the more relevant factor. Items in conditions B and C in Table 3 would provide us with the necessary test. If items on which ditone sequence frequency mismatches word frequency (condition B) behave like items in condition D, and items on which ditone sequence frequency matches word frequency (condition C) behave like items in condition A, this would provide support towards the idea that ditone sequence frequency underlies the role of tone in lexical selection. Conversely, if items on which tone bigram frequency mismatches word frequency (condition C) behave like items condition D, and items on which bigram frequency matches word frequency (condition B) behave like items in condition A, this

would provide support for bigram frequencies underlying the role of tone in lexical selection. This part of the experiment was exploratory since we did not have access to tone bigram frequencies during stimulus construction.

1. Participants and setup

47 speakers of Standard Mandarin (36 female, 11 male, mean age = 22, SD = 6, range 18-50) participated in the perception experiment. Participants were recruited through the UCLA Psychpool or using flyers on the UCLA campus. They either received course credit or were paid $10 for their participation. All participants considered Mandarin Chinese to be their native language, and had to be able to handwrite simplified characters. The ability to handwrite in characters was also taken to be an additional measure of proficiency: all but 2 participants had completed most of their education so far in mainland China (>10 years). None of them had a known history of hearing or speech impairment.

The experiment script was written in Python 2.7 (Python Software Foundation) with the PyGame module (Shinners, 2011). The experiment ran on a PC laptop connected to an external monitor, headphones and a graphic tablet. With this setup, the participant used the graphic tablet to give their responses, and had no access to keyboard or mouse. The experiment took place in the UCLA Phonetics Laboratory soundbooth. During the experiment, direct to disk audio recordings were made using a balanced-input Shure head-mounted SM10A microphone into an external XAudio box. The software PCQuirer (Scicon RandD) was used as the recording program, with a sampling rate of 22050 Hz.

2. Materials

30 pairs of test words and 60 pairs of filler words were selected. For each test pair, 3 primes were selected. Test stimuli were disyllabic, and could be interpreted as 2 different words/compounds depending on the associated tone sequences (e.g. juli could be identified as ju4li2 'distance' vs ju3li4 'to give an example', full list in Appendix II). The two possible words were semantically and orthographically unrelated to each other, and to any of the primes (i.e. no shared radicals). Tones on both syllables had to differ between the two alternatives (e.g. if one alternative was T2T3, the other could not start with T2 nor end with T3).

We needed to consider tone sequence frequency in our design. For a given target, the most frequent alternative (word frequency) could either bear the more frequent tone sequence, or not. At the time we selected these items, we did not yet have access to the tone bigram frequencies in running speech, provided by our own corpus. The tone sequence frequencies used to select the experimental stimuli were thus the ditone sequence frequencies in words obtained from the SUBTLEX-CH. Tone bigram frequencies were coded post hoc in the experimental data (Full list in Appendix III).

All items were recorded by a female adult native speaker of Beijing Mandarin. 30 test target items were created. Three different items were used to create a given target: the two disyllabic words that differed only on tone sequence (e.g. ju4li2 'distance' and ju3li4 'to give an example') and a nonce word with the same segmental content, bearing Tone 1 (high level tone) on both syllables (e.g. ju1li1). The target stimulus was created by replacing the pitch contour on the nonce word by the contour obtained by averaging the pitch on the two disyllabic alternatives. This is to limit participants' tendency to develop strategies relying on other acoustic cues (voicing, duration…) for tonal identification (Liu and Samuel, 2004). Pitch replacement was implemented in Praat (Boersma, 2001). Pitch contours were only considered on sonorant segments, manually tagged in Praat. Pitch

was measured at the same number of points for the two disyllabic words and interruptions in a contour were interpolated. The number of points where pitch was measured was determined by the duration of contour-bearing segments on the output target item: one point was acquired per millisecond. Figure 7 below shows the pitch contours involved in creating the target stimulus for 'juli':

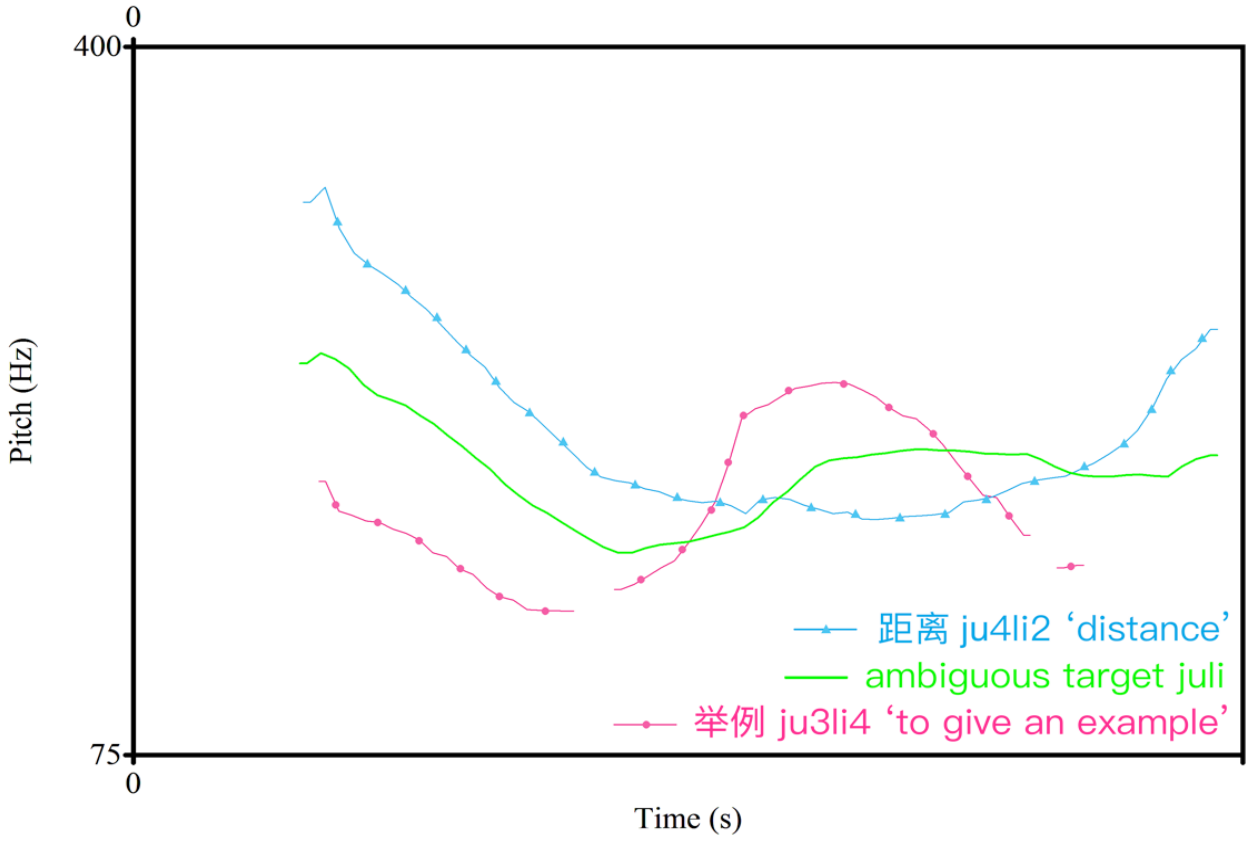

Figure 7. Pitch contours involved in creating the target stimulus for 'juli': ju4li2 'distance' in blue, ju3li4 'to give an example' in magenta, and the resulting averaged contour in green. Pitch was measured at the same number of points for the two disyllabic words and interruptions in a contour were interpolated. This allowed for normalization of durations. The green contour replaced the pitch contour on a nonce ju1li1 item, produced with level tones.

For each target, three primes were recorded. Two related primes were disyllabic words bearing the tone sequences of the possible words that the target could be identified as (e.g. shui4mian2 'sleep (noun)' prime matching the more frequent alternative ju4li2 'distance', vs. shou3juan4 'handkerchief'

prime matching the less frequent alternative ju3li4 'to give an example'). Primes were semantically and orthographically unrelated to either possible word, and matched the average of their frequency (e.g. shui4mian2 'sleep (noun)' and shou3juan4 'handkerchief' were equally frequent, and their frequency was the average of the frequency of ju4li2 'distance' and ju3li4 'to give an example'), so that the frequency of the prime did not point to either the more frequent or less frequent parts of the lexicon. We will thereafter refer to the prime matching the tone sequence of the more frequent alternative as mf prime, and the prime matching the less frequent alternative as lf prime. The unrelated prime could only bear tones that were present in neither of the two possible words (e.g. cheng1hu5 'to address' as unrelated prime for ju4li2 'distance' and ju3li4 'to give an example'). This also constrained the selection of test word pairs, as some tone combinations do not occur in disyllables (e.g. two neutral tones). For instance, a word pair such as zheng1qi4 'steam' – zheng3qi2 'organized' could not be used, since the only tone combination left for an unrelated prime would be 55, a sequence of two neutral tones.

Each participant only heard each target with one of the primes (mf, lf or unrelated); this was counterbalanced between participants.

60 filler trials consisted of disyllabic words where segmental information alone sufficed to identify the word. Filler words were chosen in pairs that were segmentally, semantically, orthographically unrelated. The word frequency distribution of the filler pairs was matched to that of the test word pairs. In order for the filler stimuli not to stand out, they were also subjected to pitch manipulation. Pitch on the two words was averaged and made to replace the contour on one of the words. A manipulated version was made for each word in a pair. For any given filler pair, a given participant heard the non-manipulated first word in place of prime, and the manipulated second word in place

of target. The manipulated word in each pair was randomized across participants. In total, each participant heard 90 trials. Filler trials were not analysed.

3. Procedure

In each trial, participants were asked to listen to pairs of words (prime + target) and to identify the second word in each pair. They were told that the second word was not pronounced clearly. Once they decided on the identity of the target word, they were asked to click down as fast as they could on the graphic tablet and to write that word down, then say it aloud. The double response was to ensure correct encoding of the response. Handwritten responses were necessary to bypass lexical prediction that keyboard answers would provide. The most frequent typewritten input system for Chinese characters uses segmental pinyin input and prompts the user to choose between the possible alternatives: this would alert our participants immediately that a given item could be ambiguous. Audio-only responses, or typewritten pinyin with tone numbers also did not suit our purpose: if participants made an ad hoc compound whose tones matched one of the predicted responses (e.g. dong1ji4 as 'east' + 'to send' rather than the predicted 'winter season'), the frequency of that item (presumably null as a word) could not be taken into account. However, audio responses were useful in case the written response was hard to decipher, or if the participant forgot how to write one of the characters (if so, they were told to write what they remembered, add pinyin with tone, and say in the recording what word they meant – paraphrases. There were 9 such cases.)

Inter-stimulus interval (ISI) between offset of prime and onset of target was 50 ms. Order of the trials was randomized, with the experiment starting with at least 5 filler trials. The experiment lasted between 30 and 40 minutes (see Figure 8 on the following page for the schematic of one trial).

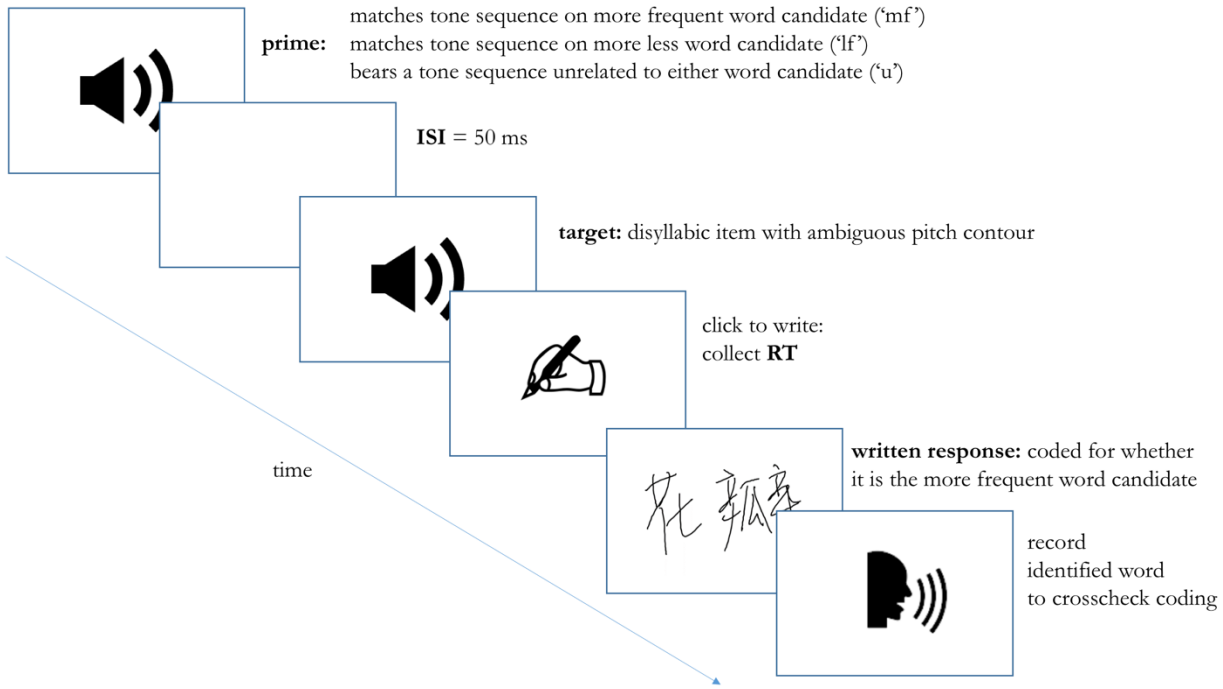

Figure 8. Event sequence for one trial in the word recognition experiment. The entire experiment consisted of 90 trials, lasting between 30 and 40 minutes.

Written responses were coded manually by a native speaker of Mandarin. Recording tracks were used to crosscheck in case the participants' response was hard to read or a character was forgotten, and to ensure that the written responses corresponded to the identified tones.

4. Data analysis

Exclusion criteria

Participants

5 participants were excluded from the analysis (1 technical issue, 1 did not follow instructions, 3 gave more than 8 responses that matched neither of the two possible word candidates - 2 standard deviations from the mean among participants (4)). Data from 42 participants was retained.

Items

To determine whether to exclude data from specific items, we considered data from all priming conditions: if the related primes could make available responses that had not been considered in the unrelated condition (for example, only more frequent word and 'other' answers in the unrelated condition, but less frequent word answered in one of the priming conditions), then we would be ignoring this difference if we had excluded these items based on the unrelated condition only. This also minimized datapoint exclusions.

Data from 6 test pairs were excluded. This was either because: the number of responses matching neither word candidate exceeded 2 standard deviations from the mean among items (1 pair)[1], or participants only ever responded with one of the possible candidates in all conditions (5 items). This could be because the audio stimulus was not ambiguous enough or because one of the word alternatives is much more prominent (word frequency or tone frequency) than the other. Thus, results from these items could not be clearly interpreted.

Data from 24 items were retained. See Appendix III for detailed repartition in frequency conditions.

Trials

Test trials where the response was not one of the two predicted word alternatives were excluded (13.1% of responses).

Although we collected reaction times as well, we did not analyse them. We asked participants to click down with the pen as soon as they had identified the 'unclear': the gesture involved more effort than

[1] Item 12: 'yingke', which we expected to be recognized as 24 'to greet' or 42 'hard shell', was also potentially the name of a recent popular smartphone application (44 映客, launched October 2015), which we were not aware of at the time of creating the stimuli.

the usual keypress used to acquire reaction times. This was to limit the amount of hardware that participants had to interact with, making them use only the pen and tablet system. Because of the orthographic requirement, we also suspected that participants might tend to click only when they were sure how to write the identified word rather than as soon as they decided on a word.

Mixed-effects regressions

We analysed our data using multilevel logistic mixed-effects regression models in Stata 13, using the me (mixed effects) set of functions. The dependent variable examined was whether for any given trial the response to an ambiguous item was the most frequent word alternative.

5. Results and discussion

Frequency match condition

We examined whether priming with a tone pattern matching the more frequent word alternative (***mf***) or matching the other, less frequent alternative (***lf***) made participants more or less likely to respond with the more frequent candidate than when they were primed with a tone pattern unrelated to either candidate (***u***). For items where the more frequent word candidate also bore the more frequent tone pattern (according to both ditone and bigram frequency; condition A), that candidate is probably very salient during lexical access. So we expected the least amount of priming in this condition. Figure 9 on page 37 shows the percentage of cases where participants answered with the more frequent word with different prime types and frequency conditions. We modeled our results using a mixed-effects logistic regression including prime type as fixed effect, and participant (SUBJ) and ITEM as crossed random effects (random slopes and intercepts). Participants were overall likely to answer with the more frequent word alternative, and priming conditions did not significantly affect this preference for the more frequent alternative (68% of more frequent word responses

across all priming conditions, above chance at 50%). Table 4 below summarizes the model coefficients. This is most likely due to ceiling effects, and we cannot distinguish the role of word and tone frequency for these items.

Table 4. Fixed effect coefficients in a model fitted to response, in cases where the more frequent word candidate bears the more frequent tone sequence (according to both ditone and bigram measures)

		Estimate	**Std. Error**	***z***	**Pr> |z|**
Intercept (unrelated prime)		0.20	0.25	0.82	0.409
Prime	lf	-0.01	0.33	-0.02	0.9810
	mf	-0.13	0.34	-0.37	0.7120

Frequency mismatch condition

A more informative case is that of items where the more frequent word alternative bears the less frequent tone pattern (as defined by both ditone and bigram frequency: condition D). If tone sequence frequency makes a word more or less salient during processing, this mismatch should reduce the saliency of the more frequent word. In this case, participants might be more sensitive to priming with tone patterns.

For these items, participants were still quite likely to answer with the more frequent word candidate in the unrelated prime condition (59% of more frequent word responses vs 68% in the frequency match condition - A). However, when presented with a prime matching the tone pattern of the more frequent word, participants were *less* likely to answer with that word ($p < .005$). Though non-significant, we also observed the complementary trend for cases where participants were primed with the tone pattern of the less frequent word: this seemed to increase their likelihood to answer the other word candidate (see Figure 9 on the following page, full coefficients in Table 5 on page 38). We think this trend would reach significance with more items and/or subjects.

We have shown earlier that in the lexicon, tone is more informative on disyllables than monosyllables. When compared to condition A, the results from condition D suggest that tone sequence frequency plays a separate role from word frequency and interferes with word frequency information during the processing of disyllables. This interference makes items where the frequencies of words and tone sequences mismatch sensitive to priming effects, but is not strong enough to influence the likelihood of choosing the more frequent word candidate when word and tone sequence frequencies match.

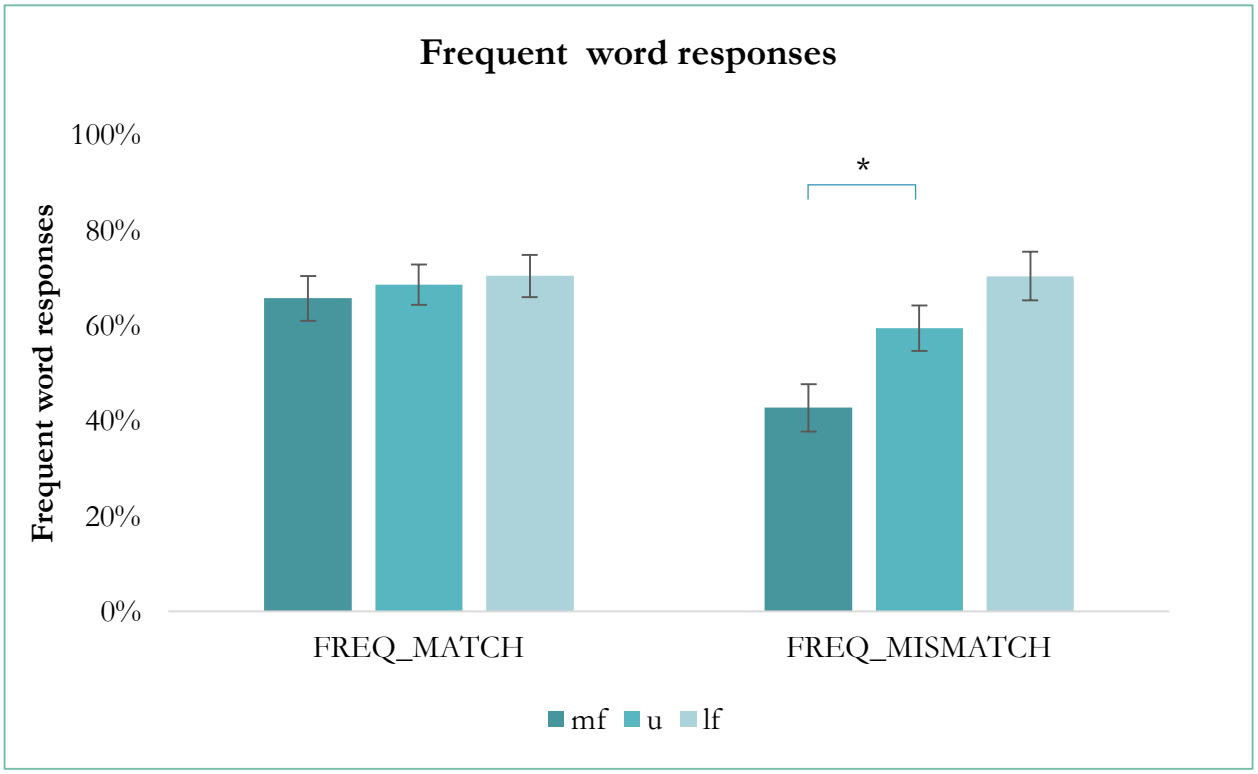

Figure 9. Frequent word responses. Priming conditions are: prime matches tone sequence on the higher frequency word alternative (***mf***), prime bears tone sequence unrelated to either word alternative (***u***), prime matches tone sequence on the lower frequency alternative (***lf***). Frequency conditions are: the more frequent word alternative bears the more frequent tone pattern according to both ditone and bigram definitions (FREQ_MATCH), the more frequent word alternative bears the less frequent tone pattern according to both ditone and bigram definitions (FREQ_MISMATCH). Error bars are standard error to the mean.

Table 5. Fixed effect coefficients in a model fitted to response, in cases where the more frequent word candidate bears the less frequent tone sequence (according to both ditone and bigram measures)

		Estimate	**Std. Error**	***z***	**Pr> |z|**	
Intercept (unrelated prime)		0.86	0.28	3.10	0.002	*
Prime	lf	0.67	0.37	1.83	0.067	
	mf	-0.74	0.33	-2.23	0.026	*

Note that the effect of priming with tone sequences, when present, is inhibitory. Inhibitory priming of tone has been reported previously for Mandarin monosyllables as well (Poss et al, 2008). One possible explanation for this phenomenon in our data would be that participants contrasted the clear tone sequence on the prime to the acoustically modified, ambiguous tone sequence on the target test item. For example, if a participant hears a clear token of shui4mian2 'sleep (noun)' before an ambiguous token 'juli', which bears a different pitch contour, they might contrast them acoustically and conclude that the ambiguous contour is not likely to be an instance of 42, since it is a much worse instance than the clear 42 just heard. Therefore, they may decide that the ambiguous word must not be ju4li2 'distance', but rather the less frequent alternative ju3li4 'to give an example'. Direct acoustic comparisons by listeners, of the sort discussed here, are typically observed at short ISIs and when the token-to-token variability is limited, as is the case here because we used single talker stimuli.

Comparing ditone sequence frequency and tone bigram frequency

The cases analysed previously do not allow us to distinguish between our two alternative measures of tone sequence frequency: either the **more** frequent word alternative had the **more** frequent tone pattern according to both ditone and bigram definitions, or the **more** frequent word alternative had

the **less** frequent tone pattern according to both ditone and bigram definitions. In the table below (Table 3, repeated for convenience), they correspond respectively to cases A and D.

Table 3 (repeated). Possible frequency conditions. For condition B, the more frequent word (e.g. zaofan 34 'breakfast' is more frequent than 43 'to revolt') bears the more frequent tone pattern (frequency match) according to ditone frequency (34 more frequent than 43 on disyllabic words), but this pattern is the less frequent alternative (frequency mismatch) according to bigram frequency (in running speech, 43 is more frequent sequence than 34).

A	**More frequent ditone (match)** More frequent bigram (match)	***B***	Less frequent ditone (mismatch) More frequent bigram (match)
C	**More frequent ditone (match)** Less frequent bigram (mismatch)	***D***	Less frequent ditone (mismatch) Less frequent bigram (mismatch)

Items from conditions B and C, in which ditone and bigram frequency disagree in whether tone sequence frequency matches or mismatches word frequency (for instance, items in condition B would be cases of frequency match for ditone, but mismatch for bigram), would allow us to determine which measure is more relevant for characterizing the influence of tone sequence frequency during lexical processing. Recall that we saw no priming in Condition A, where the more frequent word alternative also has the more frequent tone frequency. However, when word frequency and tone frequency mismatch, in Condition D, priming inhibits the word candidate whose tone pattern is identical to that the prime. In this section we investigate whether a mismatch in ditone or bigram frequency alone (Condition B or C) shows inhibitory priming of the sort observed in Condition D. If frequency mismatches in ditone alone influence lexical processing, then we should see inhibitory priming in condition B, but not C. If frequency mismatches in bigram frequency alone influence lexical processing, then we should see inhibitory priming in condition C, but not B.

We had few items in these two categories: 5 in condition B (frequency mismatch for ditone, match for bigram) and 5 in condition C (frequency match for ditone, mismatch for bigram). Figure 10 below shows the percentage of cases where participants answered with the more frequent word.

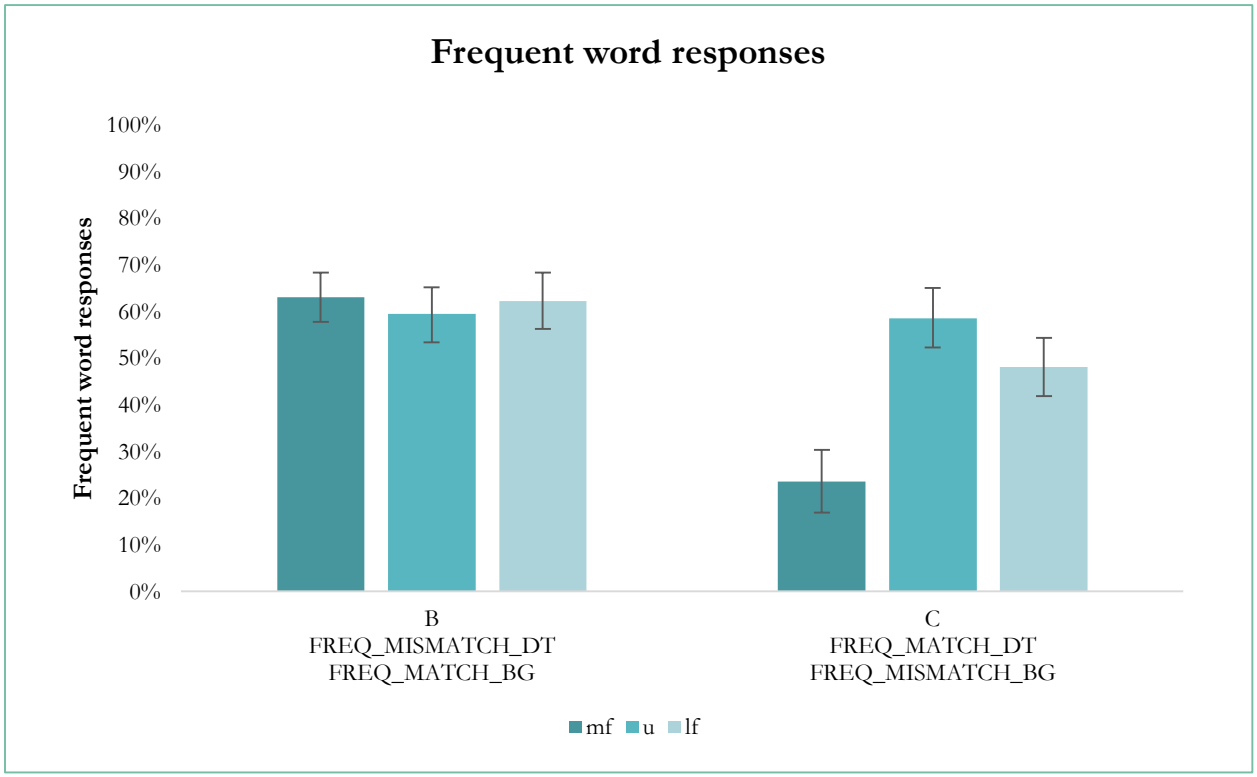

Figure 10. Frequent word responses. Priming conditions are: prime matches tone sequence on the higher frequency word alternative (mf), prime bears tone sequence unrelated to either word alternative (u), prime matches tone sequence on the lower frequency alternative (lf). Frequency conditions are: the more frequent word bears the less frequent ditone but more frequent bigram (B), and the more frequent word bears the more frequent ditone but less frequent bigram (C). Error bars are standard error to the mean.

Since there were so few data points, we will not report statistical results on these conditions. The pattern we can observe (see Table 3 repeated below for convenience) suggests however that priming conditions are inducing differences in condition C that are not present in condition B. This suggests that frequency mismatch for bigrams is most likely to account for the inhibitory priming observed in Condition D as well. If this effect generalizes to more items, this would point to tone bigram

frequency as the appropriate measure of tone sequence frequency. Specifically, it would show that Mandarin listeners' lexical selection is influenced by the likelihood of encountering a given sequence of tones in running speech, regardless of word boundaries.

Table 3 (repeated). Possible frequency conditions. For condition B, the more frequent word (e.g. zaofan 34 'breakfast' is more frequent than 43 'to revolt') bears the more frequent tone pattern (frequency match) according to ditone frequency (34 more frequent than 43 on disyllabic words), but this pattern is the less frequent alternative (frequency mismatch) according to bigram frequency (in running speech, 43 is more frequent sequence than 34). Results for conditions A and B were similar, and C and D were similar: bigram frequency seems to be the deciding factor.

A	More frequent ditone (match) **More frequent bigram (match)**	***B***	Less frequent ditone (mismatch) **More frequent bigram (match)**
C	More frequent ditone (match) **Less frequent bigram (mismatch)**	***D***	Less frequent ditone (mismatch) **Less frequent bigram (mismatch)**

Conclusion

Previous research has shown that tonal information plays a minor role during lexical access in Mandarin (Taft and Chen, 1992). However, most of the studies on processing of tone use monosyllables, which are not the most widespread word form in the modern lexicon, and might furthermore underestimate the role of tone information. To examine the role of tone in contexts where it is more informative, we conducted a set of 3 experiments.

The first corpus study on Mandarin subtitles compared the informativeness of tone in monosyllabic vs. disyllabic Mandarin words. Based on corpus counts, we showed that disyllables constitute the greater part of the lexicon, and furthermore that the contribution of segmental and tonal information is not comparable for monosyllabic and disyllabic words. Even combining segments

and tone on a monosyllable selects on average 4 homophones. In contrast, segments-only sequences can select a unique word in 78% of disyllables, a proportion equivalent to slightly over half of the lexicon. Adding tonal information on disyllables, a unique word is selected in 94% of disyllables (62% of lexicon). Thus, although overall segments contribute the greater amount of information to Mandarin lexical access, tone is much more informative on disyllables than monosyllables.

In the second experiment we constructed a spoken language corpus of semi-spontaneous spoken Mandarin to estimate alternative measures of frequency for tone sequences. Since tone is most informative at the disyllabic level, we examined frequency of tone sequences rather than individual tones. Tone sequence frequency can be computed within disyllabic words (ditone sequence frequency) from an existing corpus. However, the new spoken language corpus allowed us to obtain the likelihood of hearing a sequence of T_aT_b in running speech (across word boundaries, tone bigram frequency). We showed that ditone-based and bigram-based measures of tone sequence frequency do provide significantly different measures of tone sequence frequency.

Finally, in experiment 3, we investigated the role of tone sequences in disambiguating segmentally identical disyllabic word candidates that differed only in tone. Using a priming paradigm, we first tested whether tone sequence frequency plays a role independent from word frequency in processing disyllabic words. Results show that when word frequency and tone sequence frequency match, participants largely prefer the favoured candidate and do not show sensitivity to priming. However, when word frequency and tone sequence frequency mismatch, participants show an inhibitory priming effect and preferred the word candidate bearing a tone pattern different from the prime. These results suggest that tone sequence frequency plays a separate role from word frequency and interferes with word frequency information during the processing of disyllables. This interference makes items where the frequencies of words and tone sequences mismatch sensitive to priming

effects, but is not strong enough to influence the likelihood of choosing the more frequent word candidate when word and tone sequence frequencies match. However, these two sets of items cannot tell us which measure of tone sequence frequency is more appropriate.

Using a smaller, pilot dataset, we also examined which measure of tone sequence frequency is more appropriate to describe the role of tone in processing disyllabic words. If listeners are sensitive to frequencies specific to the lexical subset of disyllabic words, then the more relevant measure should be ditone frequency in disyllabic words, or how often a word bears a given sequence of T_aT_b. However, if listeners are only sensitive to the overall likelihood of hearing a sequence of T_aT_b in spoken Mandarin, tone bigram frequency in running speech would be more relevant. In our priming experiment, we compared items for which the two alternative measures of tone sequence frequency made opposite predictions. Participants seemed to be sensitive to mismatches between tone bigram frequency and word frequency, but not between ditone sequence frequency and word frequency. Data from more items would be needed to build a stronger argument for this case.

These three experiments allowed us to study an aspect of the role of tone in lexical selection that had not been examined previously. Previous studies focused mainly on comparing the contribution of segments and tones on monosyllables, and found that segmental information mattered more. We show that tone is inherently less informative on monosyllables, and that it is more crucial in distinguishing disyllabic words. On disyllables, the contribution of tone is mediated by frequency, and tone sequence frequency can interfere with word frequency when several lexical candidates are competing during lexical selection. Pilot results suggest that the type of tone frequency that listeners encode is the overall likelihood of encountering a given sequence of tones in running speech, regardless of word boundaries.

Finally, in order to obtain our bigram frequency measures, we acquired data to start building an annotated corpus of semi-spontaneous Mandarin. In the future, we plan on expanding this database and making it openly available for linguistic research.

Appendix I. Corpus recordings - conversation topic suggestions

北京路况	traffic in Beijing
最近听说的趣事	interesting anecdote(s) that you heard about recently
常用的手机应用	phone app(s) that you often use
换手机	changing cell phone
想去的地方	place(s) you want to go
学中文的老外	foreigners learning Chinese
烧饭/外卖	cooking/buying take out
喜欢的餐馆	restaurant(s) that you like
电脑故障	computer trouble(s)
网络游戏	online gaming
拍照	taking pictures
有意思的电影/连续剧	interesting movie(s)/series
爱看/不爱看的电视节目	TV show(s) that you like/don't like
喜欢的明星	celebrity(/ies) that you like
听音乐	listening to music

爱看的书	book(s) that you like
养宠物	pet(s)
旅游	travelling
买房/租房	buying/renting
体育活动	sports
上学时喜欢的课	favourite class(es) when you were in school
特殊经历	special experience(s)
工作/同事	work/colleagues
应酬	work-related social occasions
加入您能选任何工作	if you could have any job you wanted
带孩子	raising children
孩子上学	children's schooling
家人	family members
购物经验	shopping experience(s)
喜欢的品牌/产品	brand(s)/product(s) you like
高兴的事情	thing(s) that make you happy

北京物价	price of living in Beijing
坐地铁	taking the subway
这几天的天气	recent weather
化妆品	cosmetics
广告电话	ad phone calls
微博还是微信？	Weibo or WeChat? (popular social media in China)
用过赶集网吗？	Have you ever used Ganji? (Craigslist equivalent in China)
健康食品/补品	nutrition/health foods
保持健康	staying healthy

Appendix II. Experimental stimuli – test items and primes

pinyin	reading1	T1	prime1	reading2	T2	prime2	unrelated prime	Tu
ju li	距离 - distance	42	zhi yu - 至于 - concerning	举例 - to give an example	34	wu ye - 午夜 - midnight	cheng hu - 称呼 - to address	15
mo li	魔力 - magic	24	ji shi - 急事 - urgent matter	茉莉 - jasmine	45	dou zi - 豆子 - bean	fang kuai - 方块 - square	14
hua xue	化学 - chemistry	42	bao chou - 报仇 - revenge	滑雪 - to ski	23	mei zhun - 没准 - perhaps	da qiu - 打球 - to play ball	32
hu li	狐狸 - fox	25	mo hu - 模糊 - blurry	护理 - to nurse	43	ju chang - 剧场 - theater	tan zi - 摊子 - stand (shop)	15
bai ling	百灵 - lark	32	pao ti - 跑题 - off topic	白领 - white collar	23	tian ye - 田野 - field	kan jian - 坎肩 - cardigan	31
dong ji	冬季 - winter	14	ying cun - 英寸 - inch	动机 - motive	41	dou zheng - 斗争 - struggle	liu li - 琉璃 - glaze	22
zao fan	早饭 - breakfast	34	jie jiu - 解救 - to rescue	造反 - to revolt	43	chu zhi - 处置 - punishment	zha dan - 炸弹 - bomb	44
xue li	学历 - education	24	jue jing - 绝境 - desperate situation	雪梨 - pear	32	jian jie - 简洁 - concise	zhi yuan - 职员 - employee	22
mi li	米粒 - rice grain	34	fan pai - 反派 - opposition	迷离 - mysterious	22	zhe tang - 蔗糖 - cane sugar	lun tai - 轮胎 - wheel	21
li ke	立刻 - immediately	44	bian hu - 辩护 - to defend	理科 - scientific track	31	sun shi - 损失 - loss	zeng jia - 增加 - to add	11
hua ti	话题 - topic	42	jie chu - 解除 - to dispell	滑梯 - toboggan	21	ping gu - 评估 - estimate	bi fang - 比方 - comparison	35
ying ke	迎客 - to greet	24	ming yan - 明艳 - colorful	硬壳 - solid shell	42	bao shi - 报时 - to announce time	ping an - 平安 - safe	21
shui jiao	睡觉 - to sleep	44	di yu - 地狱 - hell	水饺 - dumpling	33	bi ci - 彼此 - eachother	dian liang - 掂量 - to weigh	15

pinyin	reading1	T1	prime1	reading2	T2	prime2	unrelated prime	Tu
bai shi	拜师 - to seek teaching	41	wai bin - 外宾 - visitor	摆饰 - decoration	35	zao zi - 枣子 - date (fruit)	fei xiang - 飞翔 - to fly	12
shou ji	手机 - cellphone	31	wan can - 晚餐 - dinner	收集 - to collect	12	gong ping - 公平 - fair (justice)	lai yuan - 来源 - source	22
ban zou	搬走 - move away	13	qing xi - 清洗 - to clean	伴奏 - accompaniment (music)	44	fang da - 放大 - to enlarge	gan jue - 感觉 - feeling	32
zhu li	主力 - main force	34	wei zao - 伪造 - to forge	助理 - assistant	43	wei ci - 为此 - therefore	shou gong - 手工 - handicraft	31
mao xian	冒险 - adventure	43	hou guo - 后果 - consequence	毛线 - wool	24	chao xiao - 嘲笑 - to mock	xiang yan - 香烟 - cigarette	11
quan jia	全家 - all family	21	de chu - 得出 - to conclude	劝架 - to pacify	44	yue guo - 越过 - to bypass	hua sheng - 花生 - peanut	11
bu jiu	不久 - shortly	43	po chan - 破产 - bankrupt	补救 - to salvage	34	ling dai - 领带 - necktie	ping jun - 平均 - average	21
wei qi	围棋 - go (game)	22	wan man - 完满 - satisfying	尾气 - exhaust fumes	34	bai shu - 柏树 - cypress	shan po - 山坡 - hillside	11
an hao	安好 - safe and sound	13	biao yu - 标语 - slogan	暗号 - code	44	bi ding - 必定 - certainly	nian ling - 年龄 - age	22
chang di	长笛 - flute	22	you ju - 邮局 - post office	场地 - terrain	34	hai lang - 海浪 - wave	zui ba - 嘴巴 - mouth	35
chong feng	冲锋 - attack	11	bing chuan - 冰川 - glacier	重逢 - to reunite	22	shi tang - 食堂 - cafeteria	kai che - 开车 - to drive	11
da ting	打听 - to inquire	35	zhen tou - 枕头 - pillow	大厅 - hall	41	ri qi - 日期 - date (time)	gang qin - 弹琴 - piano	12
dian xin	点心 - dessert	35	zhong zi - 种子 - seed	电信 - telecommunications	44	te se - 特色 - specialty	ban ji - 班机 - flight (airline)	11
fei wu	飞舞 - float	13	bin guan - 宾馆 - hotel	废物 - refuse	44	zheng jian - 证件 - papers (identity)	gan lu - 赶路 - to hurry	34

feng xian	奉献 - to offer	44	san bu - 散步 - to stroll	风险 - risk	13	chu ban - 出版 - to publish	yi zhi - 移植 - to transplant	22
gaogui	搞鬼 - to play a prank	33	shui zhun - 水准 - criterion	高贵 - noble	14	qi xian - 期限 - deadline	pai zi - 牌子 - brand	25
gao su	高速 - high speed	14	gong zuo - 工作 - to work	告诉 - to tell	45	piao liang - 漂亮 - pretty	he cheng - 合成 - synthetic	22

Appendix III. Categorization of items in frequency conditions

Items excluded as outliers are shown in italics.

A More frequent ditone More frequent bigram	***B*** Less frequent ditone More frequent bigram
moli	juli
huaxue	baishi
huli	zhuli
xueli	quanjia
like	bujiu
huati	
anhao	*maoxian*
feiwu	
weiqi	
C More frequent ditone Less frequent bigram	***D*** Less frequent ditone Less frequent bigram
zaofan	bailing
shuijiao	dongji
changdi	mili
chongfeng	shouji
	banzou
gaosu	dianxin
yingke	fengxian
	dating
	gaogui

References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Brown-Schmidt, S., and Canseco-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(2), 103-135.

Cai Q, Brysbaert M (2010) SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE* 5(6): e10729. doi:10.1371/journal.pone.0010729.

Chao, Y. (1948). *Language and Symbolic Systems*. Oxford University Press, Oxford, UK.

Chen, Y., and Xu, Y. (2006). Production of weak elements in speech–Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica*, *63*(1), 47-75.

Cutler, A., and Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics*, 59, 165-179

Denisowski, P. A. (1997). CC-CEDICT public-domain Chinese-English dictionary. Retrieved from https://cc-cedict.org

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(02), 143-188.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. LOT, Utrecht.

Fox, R. A., and Unkefer, J. (1985). The Effect of Lexical Status on the Perception of Tone/成词状况对声调感知的影响. *Journal of Chinese Linguistics*, 69-90.

Gandour, J. (1983). Tone perception in far eastern languages. *Journal of Phonetics*, 11, 149 – 175.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 110 – 125.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (Vol. 6). Cambridge University Press.

Hu, J., Gao, S., Ma, W., and Yao, D. (2012). Dissociation of tone and vowel processing in Mandarin idioms. *Psychophysiology*, 49(9), 1179-1190.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Lee, C. Y. (2007). Does horse activate mother? Processing lexical tone in form priming. *Language and Speech*, 50(1), 101-123.

Liu, S., and Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*, 47(2), 109-138.

Marslen-Wilson, W. D. (1984). *Parallel processes in lexical access*. Unpublished manuscript, University of Cambridge.

McEnery, T., and Xiao, R. (2008). CALLHOME Mandarin Chinese Transcripts - XML version LDC2008T17. Philadelphia: Linguistic Data Consortium

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... and Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Oh, Y. M., Pellegrino, F., Coupé, C., and Marsico, E. (2013, August). Cross-language comparison of functional load for vowels, consonants, and tones. In *Interspeech* (pp. 3032-3036).

Poss, N., Hung, T. H., and Will, U. (2008). The effects of tonal information on lexical activation in Mandarin. In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)* (Vol. 1, pp. 205-211).

Python Software Foundation. *Python Language Reference, version 2.7*. Available at http://www.python.org

Repp, B. H., and Lin, H. B. (1990). Integration of segmental and tonal information in speech perception: A cross-linguistic study. *The Journal of the Acoustical Society of America*, 87(S1), S46-S46.

Sereno, J. A., and Lee, H. (2015). The contribution of segmental and tonal information in Mandarin spoken word processing. *Language and Speech*, 58(2), 131-151.

Shinners, P. (2011). *PyGame - Python Game Development*. Retrieved from http://www.pygame.org

Shuai, L., Li, B., and Gong, T. (2012). Priming Effects of Tones and Segments in Lexical Processing in Mandarin. In *6th International Conference on Speech Prosody*.

StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Surendran, D., and Levow, G. A. (2004). The functional load of tone in Mandarin is as high as that of vowels. In *Speech Prosody 2004, International Conference*.

Surendran, D., and Niyogi, P. (2003). *Measuring the functional load of phonological contrasts*. Unpublished manuscript, Chicago, IL.

Taft, M., and Chen, H. C. (1992). Judging homophony in Chinese: The influence of tones. *Advances in psychology*, *90*, 151-172.

Whalen, D. H., and Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*(1), 25-47.