Computational Linguistics and Chinese Language Processing
Vol. 18, No. 3, September 2013, pp. 81-106

Tones of Reduced T1-T4 Mandarin Disyllables

Shu-Chuan Tseng*, Alexander Soemer*, and Tzu-Lun Lee*

Abstract

The lexical meaning of Chinese words is determined by syllables and lexical tones. Phonologically, there are four full tones. Empirically, however, it remains a puzzle how tones are recognized when they are reduced in natural speech. This article presents three studies on tones of reduced disyllables: (1) a corpus study on disyllabic reduction, (2) two tone categorical identification experiments on fully pronounced and reduced disyllables, and (3) an analysis of word identification responses of two disyllables. Utilizing a segment-aligned corpus, disyllables were classified by ear into four degrees of contraction (from none to full), *i.e.*, where a disyllable is gradiently reduced towards one syllable. The results suggested that the onset of the second syllable was most likely to be shortened or deleted. For studying the lexical effect of tones, a Ganong-style word bias experiment was conducted on T1-T4 continua of three T1-T4 disyllables. Results of the fully pronounced stimuli confirmed that the lexical status of the disyllables affected the tone classification of F0 contours along a continuum from T1 to T4, showing distinct differences of tone identification in real words and nonwords. Then, this effect disappeared when the onset of the second syllable was removed to simulate a partly reduced disyllable. Insufficient segmental information seemed to deactivate the word-nonword contrast, *i.e.* lexical status seemed to override any acoustic information available. Tones tended to be recognized as those from a real word throughout the continua. Finally, responses to two T1-T4 disyllables from the identification experiment done by Tseng & Lee (2010) were re-analyzed. The results suggested that reduction degree, F0 shapes, word unit type, and exposure frequency seemed to play a role in the recognition of words and tones.

Keywords: Taiwan Mandarin, Disyllabic Words, Tone Perception, Reduced Speech.

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-mail: {tsengsc, soemer }@gate.sinica.edu.tw; tzulun@gmail.com

1. Introduction

How tones are recognized in natural conversation remains a puzzle, partly because a number of other factors are acting simultaneously. Some of these factors, for instance pragmatic contextual information and the degree of speech clarity, can be directly observed and manipulated when researchers analyze natural speech. The speaker- and recipient-related individuality, such as their prior experiences about the world, prior experiences about their acquired language, and their speech competence, also are important but difficult to control. This means that the substantial construction of the mental lexicon of a speaker or a group of speakers of a language cannot be investigated in-depth unless we can cope with the variability derived from discrepancies and similarities among individuals, language communities, and their concurrent interaction in speech communication. Working towards this research goal, we made an initial but significant move in this article by beginning to examine the issue of tone recognition in reduced disyllables in Mandarin Chinese. The production and the perception of lexical tones closely and dynamically correlate with the context of lexical constituency. Thus, we first focused on disyllabic words by examining their phonetic forms. Selected disyllabic content words extracted from a corpus of conversational speech were analyzed to obtain typical reduction patterns in terms of reduction degree. The reduction pattern then was applied in our later tone categorical identification experiments to simulate the contrast of fully pronounced and reduced speech. As disyllables with an internal word boundary may not have the same reduction pattern as disyllabic words, we further re-analyzed the responses of a previously conducted word identification experiment, focusing on one disyllabic word and one monosyllabic disyllable to study the role of tone as modulated by the degree of reduction.

1.1 Processing of Spoken Words

The most commonly used speech form is daily face-to-face conversation. The consensus in building natural speech corpora is the belief that we should look at realistic data for studying linguistic patterns and human behavior. The fine (and complex) details in natural speech cannot be thoroughly investigated until annotated speech corpora have been made available for different languages, for instance the Kiel Corpus of Spontaneous Speech for German (IPDS, 1995), the Chinese Annotated Corpus of Spontaneous Speech (Li *et al*., 2000), the Spoken Dutch Corpus (Goddijn & Binnenpoorte, 2003), the Buckeye Corpus of Conversational Speech (Pitt *et al*., 2006), the Corpus of Interactional Data for French (Bertrand *et al*., 2008), the Corpus of Spontaneous Japanese (Maekawa, 2009), and the Taiwan Mandarin Conversational Corpus (Tseng, 2013). Variation of spoken words has been studied since then on different levels, including segmental, word category, word frequency, and prosodic position. Corpus-based studies are required to provide quantitative empirical description of reduction in natural speech. For instance, using the Buckeye Corpus, Jurafsky *et*

al. (2001) studied the effect of lexical frequency and local contextual information on the probability of reduction for function and content words. They suggested that this shortening process was not necessarily correlated with vowel reduction, but probably with the lexical frequency. Similarly, Meunier & Espesser (2011) analyzed the Corpus of Interactional Data for French. Although not directly affected by the lexical frequency effect, they found that vowel reduction, including duration shortening and vowel centralization, occurred more often in monosyllabic function words than in monosyllabic content words. Both studies addressed the importance of lexical frequency on the surface forms of reduced word in natural speech. This leaves the question of how reduced spoken words are recognized by humans.

A number of psycholinguistic models have been proposed to explain how humans perform the task of spoken word recognition and how words are stored in the mental lexicon, including the phonetic-acoustic form and the associated higher level lexical information. The Cohort Theory, the Interactive-Activation Model (TRACE model), and the Neighborhood Activation Model were the classical models accounting for abstract lexical representation and extraction (Marslen-Wilson, 1987; McClelland & Elman, 1986; Luce & Pisoni, 1998). The Cohort Theory proposes that spoken word recognition involves both early bottom-up processing mainly utilizing word-initial onsets and late top-down contextual information. The Interactive-Activation Model states that (once the sensory acoustic-phonetic input comes in) different levels of a lexical item are activated, including feature, phoneme, and word. The Neighborhood Activation Model stresses the importance of similarity neighborhood density between words and the relative effects of word frequency on the spoken word recognition. Different from these abstract form extraction models, Lacerda (1995) suggested that prototypes of word forms are stored in memory and discrimination extent is used to filter the best exemplar, which is most similar to the prototype. To some extent, the concept of best exemplar is similar to the concept of episodic memory traces, which may preserve the (most likely) surface phonetic details (Goldinger, 1996). In speech communication, we constantly encounter different types of phonetic forms of words, which we may classify into phonological variants. Thus, it is not surprising that production frequency of words has some influence on the phonological representation of pronunciation variants, as found by Ranbom & Connine (2007) in their studies on the nasal flap in English.

While pronunciation variants are normally concerned with substitution of segments, the phonetic forms of words can deviate substantially from the canonical forms in the way that segments are omitted. Nevertheless, spoken words with a number of deleted segments still can be recognized easily, given proper contextual information. Frauenfelder & Tyler (1987) emphasized the importance of context in empirical practice of spoken language, in addition to the five typical, sequential lexical processing phases they proposed: initial lexical contact, activation, selection, word recognition, and lexical access. A number of previous studies have

made similar proposals. Words in semantically coherent sentences were more accurately recognized than those in isolation, and words heard in context often are recognized long before their full acoustic signal has been delivered (Grant & Seitz, 2000; Grosjean, 1980; Marslen-Wilson, 1987). Nevertheless, words with omitted segments have been studied only recently. Highly reduced words were well recognizable only when presented in their original context with semantic and syntactic information. Without context or only with limited contextual information of adjacent syllables, they cannot be properly recognized (Ernestus *et al.*, 2002; Tseng & Lee, 2010). Thus, questions are raised. Do the same sequential lexical processing phases apply to the recognition process of reduced words without context (Frauenfelder & Tyler, 1987)? Will we observe a lexical effect of tones in reduced words without contextual information? In this study, we aim to study the reduction patterns of Mandarin words in natural speech and how tones in reduced words are recognized.

1.2 Tones

Lexical tones in Mandarin Chinese constitute an abstract, contrasting phonological system of four full tones and a neutral tone (Ho, 1996; Duanmu, 2000). The four full tones include the high level tone (T1), the mid-rising tone (T2), the dipping tone (T3), and the high falling tone (T4). The neutral tone is normally noted as T5. The lexical meaning of Chinese words is determined by the syllable and tone information simultaneously. For decades, the issue of how lexical tones are produced and recognized has attracted tremendous attention and interest from linguists and psycholinguists. Tones are often illustrated by means of fundamental frequency contour (F0). F0 contour is a kind of acoustic information transmitted through the air to the ears of the listeners. Canonical forms of lexical tones in monosyllables show similar F0 contour shapes (Xu, 1997: 67; Lai & Zhang, 2008: 184). A clear high level F0 contour is observed for T1. T4 starts higher than T1 with a falling contour. T2 and T3 start lower in pitch height, with a rising and dipping contour, respectively. In addition to studies on monosyllables produced in isolation, contextual effects of tones have also been investigated. Very detailed phonetic cues have been studied or used for manipulating the stimulus tokens, such as the onset and the offset F0 values of syllables and those in the adjacent syllables, as well as the degrees of rise and fall of the F0 contours (Lin & Wang, 1984; Xu, 1997; Xu, 2004; Cutler & Chen, 1997). A number of different paradigms have been applied to test how tones are perceived or recognized (Lai & Zhang, 2008; Lee, 2007; Lee *et al.*, 2008; Hallé *et al.*, 2004; Ye & Connine, 1999; Fox & Unkefer, 1985). Testing monosyllables with four lexical tones in the gating experiment of Lai & Zhang (2008), T1 and T4 are confusing to the listeners in the first 40 to 80 ms after onset. After a period of 80 ms, the recognition turns to be correct in cases of both T1 and T4. The vowel and tone monitoring tasks conducted by Ye & Connine (1999) provided support for the notion that tones are activated differently (from vowels) when

the listeners hear the syllable-tone stimulus tokens in isolation or in context. To more concretely involve the semantic association of experiment stimuli, Lee (2007) and Lee *et al*. (2008) took into account semantic and acoustic cues in their form priming and identification experiments. Their results suggested that tonal information can be used to reduce activated candidates and can be compensated for when segmental information is insufficient. To examine the lexical status effect of tones in an implicit way, the Ganong paradigm (Ganong, 1980) was adopted by Fox & Unkefer (1985). When using the Ganong paradigm, listeners are given a standard categorical perception identification task, *i.e*., they are played a series of phonetic stimuli that vary from one phonemic category to another in a gradient way and must identify each stimulus as belonging to one of these two phonemic categories. One end of the continuum is a real word and the other is not, usually finding a response preference for the phonemic category that forms a real word (*e.g*., in a tash-dash continuum, English listeners will identify the /d/ in more stimuli than /t/, since "dash" is a real word but "tash" is not). Fox & Unkefer (1985) tested the lexical status effect of tones by having native Chinese and non-native English subjects identify T1 or T2 in the stimulus tokens of T1-T2 continua of Mandarin Chinese monosyllabic words and nonwords. Differences were observed between the two groups of subjects. Nevertheless, no differences were found between the nonword/nonword continuum and the other continua involving real words. Hallé *et al*. (2004) conducted a tone identification task on tone continua of T1-T2, T2-T4, and T3-T4 monosyllables to native Taiwan Mandarin and non-native French listeners. Despite observable differences between the native and non-native groups, the French listeners were also sensitive to the changes of F0 shapes in perceiving the tonal contrast to a certain degree. As we speculate on the lack of an obvious semantic association of the stimulus tokens in the above experiments, we used disyllabic words instead of monosyllabic words in our tone perception experiment to concretely enhance the link to the lexical constituency.

2. Disyllabic Words in Mandarin Chinese

2.1 Why T1-T4 Disyllabic Words?

The majority of words in modern Mandarin Chinese are either monosyllabic or disyllabic, according to the 5-million words of a textual corpus, the Sinica Balanced Corpus[1] (Chen & Huang, 1996). In terms of word tokens, mono- and disyllabic words together make up approximately 90% of the corpus. Tri- and quadra-syllabic words normally are composed of mono- or disyllabic words. **Figure 1** illustrates the frequency percentages of monosyllabic words, disyllabic words, and words of more than two syllables in the Sinica Balanced Corpus of 5 million words, in the Taiwan Mandarin Conversational Corpus of 500K transcribed

[1] Website: http://db1x.sinica.edu.tw/kiwi/mkiwi/.

words [2] (Tseng, 2013), and in the historical work *Zuozhuan* of 160K words from approximately 400 B.C. (the Academia Sinica Tagged Corpus of Old Chinese[3], Wei *et al*., 1997).

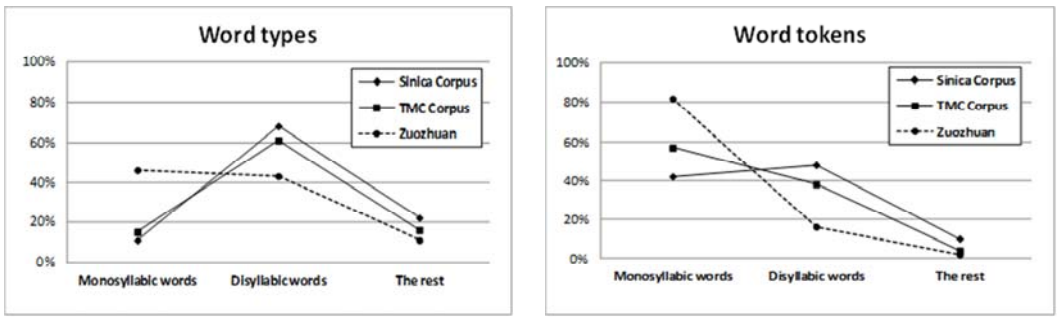

Figure 1. Word distribution in Mandarin Chinese corpora of modern texts, modern conversational speech, and ancient text.

The use of monosyllabic words descends and that of disyllabic words ascends in modern Mandarin in both the written and spoken uses. The Sinica Balanced Corpus and the Taiwan Mandarin Conversational Corpus share a similar distribution of disyllabic words, both making up 60% of the word types and 40% of the word tokens in the corpus. According to the publicly-distributed Taiwan Mandarin Spoken Wordlist derived from the Taiwan Mandarin Conversational Corpus (Tseng, 2013), the four most frequently produced disyllabic tone pairs are T4-T4, T1-T4, T2-T4, and T3-T4 with a proportion of 13.4%, 9.7%, 8.9%, and 8%, respectively. One of the reasons we used T1-T4 tone pair in our later tone perception experiment was both T1 and T4 are produced with a high pitch onset. This shared property makes it easier to manipulate the tone continua than the other tone pairs.

2.2 Using Syllable Contraction to Define the Reduction Degree of Disyllables

There should be a wide variety of reduced word forms, given different syllable structure, phonological neighborhood, lexical constituency, sentence structure, *etc*. Thus, how to represent the typical reduced word form is the first task we encounter. As we are concerned with disyllabic words, we chose syllable contraction to annotate the degree of reduction. Syllable contraction is a phenomenon of producing words of more than one syllable in a shorter way, resulting in a decrease of the number of segments and possibly also syllables. Some of the syllable mergers are predictable in Chinese phonology and dialectology (Lung, 1976; Cheng, 1985; Chung, 1997; Lien, 1997). The onset of the first syllable and the rhyme of

[2] Website: http://mmc.sinica.edu.tw.

[3] Website: http://old_chinese.ling.sinica.edu.tw/.

the second syllable make up the final form of the merger, with the rhyme of the first syllable and the onset of the second syllable omitted, which is also known as the Edge-in Theory (Chung, 1997; Hsu, 2003). For instance, the merger of two sentence-final particles *zhi55* and *hu11* in Old Chinese is represented by a new character with the pronunciation *zhu55*. This kind of lexicalized syllable merger may be due partly to speech reduction in natural speech use, as different degrees of syllable contraction including the merger are observed in production data of natural speech (Tseng, 2005). In our corpus analysis of disyllabic words, we adopted the concept of syllable contraction to annotate four different degrees of reduction.

3. Reduction of Disyllabic Words in Natural Speech

In our study of disyllabic word reduction, we examined a dataset of 3.5 hours of conversational data produced by 16 speakers to account for speaker differences and to create a more or less near-"naturalness" of the speech data. This dataset was extracted from the Taiwan Mandarin Conversational Corpus (Tseng, 2013). The transcripts of the conversations were automatically segmented and POS tagged using the CKIP (1998) system. Segmentation errors, including errors of proper nouns, idioms, constructions with numbers, directional complements, and disfluencies, were manually corrected (Tseng, 2013). To avoid highly frequent function words, as they may have specific phonetic forms due to their semantic predictability in speech communication, we analyzed disyllabic verbs and nouns. Those that contained reduplicated syllables, such as *ba4ba5* (colloquial form for "father"), and homophonous words, such as *ge4ren2*, which can mean oneself or each one, were excluded. As a result, 1,496 tokens of 50 disyllabic proper nouns and verbs that were produced at least three times in the corpus were studied (**Appendix A**).

3.1 Human Labeling of Reduction Degree in Terms of Syllable Contraction

Although previous speech production studies on syllable contraction of Taiwan Mandarin observed word forms contracted to different extents (Tseng, 2005), they lack an adequate definition for operationally annotating the instances. In our current study, we propose an annotation scheme consisting of four reduction degrees. If there is an audible syllable boundary within a disyllabic word, it is marked as **uncontracted** one. In the case of a vague syllable boundary, it is marked as **moderately contracted**. If the two nuclei of a disyllabic word are merged and it is practically impossible to separate the two syllables anymore, the boundary is regarded as **considerably contracted**. In the case of syllable merger, it is a **merged** case. Two annotators independently labeled the reduction degree of each disyllabic word following the above criteria. The inter-rater's agreement is high, Kappa = .803, $p < .01$. Forty inconsistent cases were discussed until a consensus was achieved. As a result, half of the

instances were marked as **merged** (724 occurrences). There were 171 and 300 **Moderately** and **considerably contracted** instances, followed by 301 **uncontracted** occurrences. The four annotated categories are well-balanced, which provides a good basis for the later analysis.

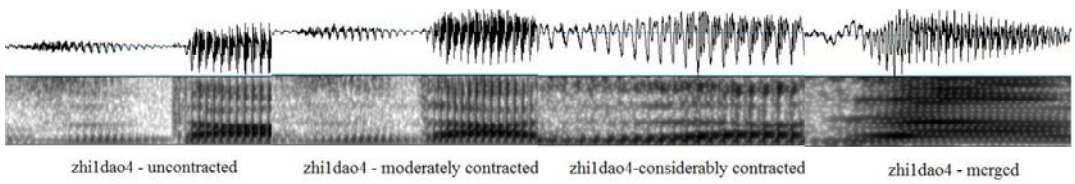

Figure 2. Four reduction degrees.

Figure 2 illustrates the spectrograms of the T1-T4 disyllabic verb *zhi1dao4* /tʂ˥ tau/ (to know) annotated with the four reduction degrees. They were all located in a prosodically medial position. The first three examples of *zhi1dao4* were spoken by a male speaker, and the merged example was spoken by a female speaker. In the uncontracted instance, the spike of /t/ is clearly observed, whereas it is not the case in the moderately contracted example. The syllabic boundary in the moderately contracted example is clear, but it is hardly visible in the considerably contracted example. In the merged case, no syllabic boundary can be observed in the spectrogram.

3.2 Segmental Reduction in Disyllabic Words

To study segment reduction in disyllabic words, we need segment-labeled data. A forced segment aligner was adopted to process the segment labeling automatically. A number of supervised and unsupervised automatic aligners have been developed to obtain segment-labeled data of realistic speech (Pitt *et al*., 2006; Scharenborg *et al*., 2010; Kuo *et al*., 2007). For our study, we trained the aligner with a human-checked, segment-labeled subset extracted from the Taiwan Mandarin Conversational Corpus with mono-phone acoustic models using the HTK toolkit (Liu *et al*., 2013). Applying this to our dataset of 1,496 tokens of disyllabic verbs and nouns, we obtained time-aligned boundaries for 7,166 segments. Please note that the forced alignment of segments was conducted independently from the labeling of reduction degree above. Given a text and a sound file, a segment aligner forcedly assigned segment boundaries to all segments in the text according to the similarity in terms of the previously trained acoustic models for each segment. That is, for the merged and contracted cases, segments that were not distinguishable by the human ear were assigned with boundaries as well. In such cases, the aligner could not find a suitable acoustic model to match the probably deleted segments, so a minimum duration of 15 ms was assigned, suggesting that it may be a case of segment substitution or deletion. Thus, we regard segments with duration of 15 ms as our candidates for deletion. In adopting this approach, instead of human judgment of

phonological variants of spoken words, we referred to the prosodic-acoustic features of the disyllabic words for our analysis of reduction degree.

The duration of all 7,166 segments was extracted via Praat (Boersma & Weenink, 2013). Please note that the coda position can only be occupied by nasals [n, ŋ] and there are only two glides [j, w]. The originally extracted segment duration was normalized by taking the z-score for each speaker, noted as **Z_dur** (Lobanov, 1971). **Table 1** summarizes the descriptive statistics of the syllabic positions of segments, including the distribution of the deletion candidates and the **Z_dur** mean duration. S1 denotes the first syllable; S2 the second.

Table 1. Segments categorized in syllabic positions.

Segment Position	Total	Deletion candidates	%	Z_dur mean
Onset_S1	1468	73	4.97%	-0.009
Glide_S1	374	34	9.09%	-0.433
Nucleus_S1	1496	273	18.25%	0.003
Coda_S1	524	229	43.70%	-0.329
Onset_S2	1340	781	58.28%	-0.484
Glide_S2	258	119	46.12%	-0.379
Nucleus_S2	1496	84	5.61%	0.720
Coda_S2	210	6	2.86%	0.062

Generalized linear models were conducted to statistically verify the contribution of the position of segment within the disyllabic words and the annotated reduction degree to the duration and the deletion likelihood of syllabic positions. Dependent variables were **Z_dur** mean and the percentage of deletion candidates. Predictors were the syllabic position of the segment and the reduction degree. The results confirmed significant effects of segment position and reduction degree on the duration and deletion percentage. Detailed results are summarized in **Table 2**.

Table 2. Summary of the Generalized Linear Models.

Dependent variable: **Z_dur** mean				Dependent variable: Deletion percentage			
	Wald χ^2	*df*	*P*		Wald χ^2	*df*	*p*
Intercept	.01	1	.92	Intercept	55.81	1	< .01
Segment position	91.04	3	< .01	Segment position	19.76	3	< .01
Reduction degree	486.01	7	< .01	Reduction degree	35.57	7	< .01

Using the uncontracted tokens as the comparison baseline, their duration was significantly longer than the other three categories ($p < .01$) and the percentage of deletion was significantly lower than the merged cases ($p < .01$), considerably contracted cases $p = .011$, and moderately contracted cases $p = .45$. To take Onset_S2 as the comparison baseline, its duration was significantly shorter than Onset_S1, Nucleus_S1, and Nucleus_S2, Coda_S2 ($p < .01$) and was shorter, but not significantly, than Coda_S1 ($p = .68$) and Glide_S2 ($p = .37$). Onset_S2 was longer than Glide_S1, but this was not statistically significant ($p = .28$). With regard to deletion percentage, that of Onset_S2 was significantly larger ($p < .01$) than Onset_S1, Glide_S1 ($p < .05$), Nucleus_S1, Nucleus_S2, and Coda_S2. The deletion percentage of Onset_S2 was larger (but not significantly) than Coda_S1 ($p = .6$) and Glide_S2 ($p = .58$).

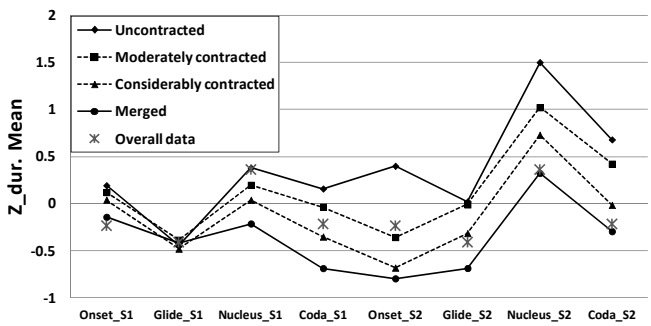

Figure 3. Temporal patterns of different reduction degrees.

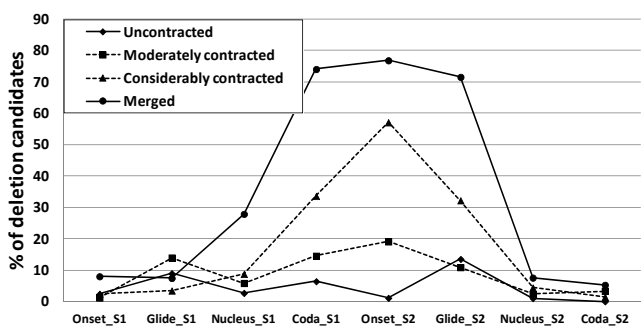

Figure 4. Distribution of deletion candidates of segments.

Figure 3 and **Figure 4** illustrate these tendencies. The asterisks mark the **Z_dur** means for each syllabic position in the overall data. Generally speaking, when a word is reduced more, the duration of segments also declines more. The onset of S2 is the shortest segment position, with an exception of the glide of S1 (Glide_S1). **Figure 3** shows a longer duration of the nucleus of S2 (Nucleus_S2) than all of the other segment positions. It became even longer

in the contracted and merged cases. A similar tendency also was observed for Onset_S1 and Coda_S2. They are both longer than the average mean durations. These are supportive evidence for the preservation of Onset_S1, Nucleus_S2, and Coda_S2, corresponding to the final form predicted by the Edge-in Theory for disyllables. As proven in the statistical results, Onset_S2 may be the most likely position to be deleted when producing the reduced disyllabic words, shown in **Figure 4**. Coda_S1, Onset_S2, and Glide_S2 contain more deletion candidates; Nucleus_S2, Coda_S2, and Onset_S1 considerably fewer. Among them, Onset_S2 contains the most deletion candidates, with significantly more than Coda_S2, Onset_S1, Nucleus_S2, and Glide_S1. The results of the deletion percentage and the duration of syllabic positions suggest that Onset_S2 may be the most likely segment to be deleted in naturally spoken, reduced disyllabic words. Please also note that, as we regarded the likelihood of segment deletion as the percentage of deleted segments in each given position separately, the syllable structure should not cause obvious artifacts on this conclusion.

Despite the widely-accepted and approved conclusions that reduced segments are more likely to be shortened, we found that changes in the duration of segments vary depending on their position in our data of disyllabic words. The results of the corpus-based segment analysis suggested that the device of reduction in disyllabic words in natural speech is systematic. Taking into account the pattern presented above, the possibly reduced forms of the CV-CV disyllabic word *zhi1dao4* (to know) may be [tʂ˩-au], [tʂau], and [au] by omitting the Onset_S2 first, then Nucleus_S1, finally also Onset_S1. In the subsequent tone categorical identification experiment, we will remove Onset_S2 to simulate the reduced version of our CV-CV disyllabic stimuli.

4. Tone Identification in T1-T4 Disyllabic Words

4.1 Data and Method

Using the Ganong paradigm for examining lexical effect of tones, the response preference should be found for the tone category that forms a real word. Nevertheless, Fox & Unkefer (1985) experimented on T1-T2 continua of monosyllabic Chinese words and nonwords without finding clear differences between nonword-nonword pairs and other pairs involving real words. This may be due to the lack of a sufficient link to semantic association of monosyllables, especially when presented in isolation. To enforce the semantic involvement in the test items, we used disyllabic instead of monosyllabic words in our study. Two T1-T4 disyllabic words *zhi1dao4* (to know) and *zhi4du4* (the system) were used as our stimulus tokens, whose T4-T4 counterparts *zhi4dao4* and *zhi1du4* are nonwords. *Zhi1dao4* was chosen, because it is the most frequently used disyllabic verb that is composed of T1 and T4 and because *zhi1dao4* and *zhi4du4* have no homophones in the Taiwan Mandarin Spoken Wordlist

(Tseng, 2013). Fox & Unkefer (1985) used DEI in their T1-T2 continua study as the nonword-nonword item. In our study, we also used *dei4* to be the nonword-nonword token. Empirically confirmed, *dei4* was not found in the Taiwan Mandarin Spoken Wordlist. As *dei4* is a non-syllable, we used it to contrast the non-word tokens with the real syllable *du4*. **Table 3** illustrates the design of the stimulus tokens.

Table 3. Stimuli design.

T1⇨T4 *zhi* continua	Contrast: Real word vs. nonword
DAO4	Real word⇨Nonword (with a real tonal syllable) *Zhi1dao4* ⇨ *zhi4dao4*
DU4	Nonword (with a real tonal syllable)⇨Real word *Zhi1du4* ⇨ *zhi4du4*
DEI4	Nonword (with a non-tonal syllable)⇨Nonword (with a non-tonal syllable) *Zhi1dei4* ⇨ *zhi4dei4*

The stimulus tokens were recorded and manipulated using Praat. *Zhi1dao4* was recorded and *zhi1* was cut out to manipulate the continuum from T1 to T4. As *zhi1* is produced with the high level T1, followed by the plosive /t/ as Onset_S2, the F0 contour of *zhi1* is slightly rising for the main portion of the entire syllable and slightly falling towards the end due to the closure of the following plosive /t/ and the high falling T4 in the second syllable. A T1-T4 continuum was synthesized by shifting the highest F0 value of *zhi1* 5 ms each time backwards while maintaining the original rising and falling slopes of the F0 contours. In this way, we increased the falling part of *zhi1*. Finally, we obtained the high-falling *zhi4*. When we concatenated the T1-T4 continua with *dao4*, *du4*, and *dei4*, we accordingly modified the pitch registers fitting those of the subsequent syllables to make each transition phase within the disyllables sound as smooth as possible. As proposed by Lin & Wang (1984), the tone recognition of a Mandarin Chinese syllable can be affected by the pitch height of the following syllable, which was taken into account when we were manipulating the tokens. At first, the tone continuum contained thirteen steps. Ten subjects were asked to judge the similarity of the thirteen steps that were presented to them in a row. As a result, we excluded five steps that most of the subjects reported as indistinguishable. Please note that the recording of the three disyllables maintained the same duration over all of the steps. While manipulating the stimulus tokens, we did not change the temporal quality of the words at all. For the first experiment, the disyllabic stimuli were fully pronounced. In the second experiment, the Onset_S2 /t/ and its transition phase to the nucleus of the second syllable were removed.

4.2 Tone Identification in Fully Pronounced Disyllables

36 subjects, 18 female and 18 male, aged between 20 and 30 were recruited. They were paid for the experiment. In the training session, the subjects heard the examples of *dian4shi4* (television) with different F0 contours contrasting with a nonword *dian1shi4* to familiarize the subjects with the procedure of the experiment. In the main session, the eight steps in each of the three T1-T4 continua *zhi1dao4 -> zhi4dao4*, *zhi1du4 -> zhi4du4*, and *zhi1dei4 -> zhi4dei4* were used as our stimuli. The acoustic properties of the first tone in each of the word-nonword, nonword-word, and nonword-nonword disyllables were the same in each step. In total, 48 trials (3 word types x 8 steps x 2 repetitions) were presented to the subjects in a randomized order and their responses were recorded using the E-Prime Software (Schneider *et al*., 2002). The task of the subjects was to decide whether the first syllable was T1 or T4 by pressing one of the two buttons marked as T1 and T4.

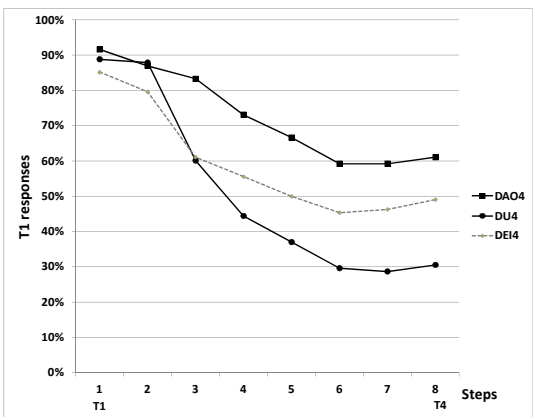

Figure 5. Tone identification responses (T1) for T1 to T4 continua in fully pronounced disyllables zhi1dao4, zhi1du4, and zhi1dei4.

A look at **Figure 5** gives the impression of a clear lexical status effect on the recognition of the first syllable. In the first two steps, the difference seems to be marginal. Starting from Step 3, DU4 and DEI4 drop drastically, compared with DAO4. The listeners still tend to hear T1, because T1 would form the real word, *zhi1dao4*. In the following five steps, the listeners continued showing this preference for the DAO4 trials. This also shows the neutral role the nonword-nonword continuum of DEI4 trials plays in the experiment. The T1 recognition rate of DEI4 is in the middle of the word-nonword DAO4 continuum and the nonword-word DU4 continuum.

We sought to confirm this subjective impression by fitting mixed-effects logistic regression models on our data (see Baayen *et al*., 2008 for a related tutorial). The basic method consists of establishing a reasonable random effects structure (which is equivalent to testing the significance of these factors with χ^2-Tests (Baayen *et al*., 2008)) before testing for

the significance of the fixed factors using an alpha level of 5%. Our base model consisted of *word type* as the only fixed factor and *step*, *subjects*, and *repetitions* as random factors. Exclusion of the random factors *step* and *subjects* in this base model led to a significant decrease in model fit (both $p < .01$), while there was no significant difference between the basic model and a model without the random factor *repetition* ($p=.48$). Thus, the factors *subjects* and *steps* were kept in the model and the factor *repetitions* was discarded. The random effects structure of this model was further extended step by step to include interactions between the fixed and random factors. The final model included a *word type* by *step* interaction ($p < .01$) and a *word type* by *subject* interaction ($p < .01$). After establishing the random effects structure, we tested the fixed effect of word type on tone recognition. The χ^2-Test showed that exclusion of the factor *word type* resulted in a significant decrease of model fit ($p < .05$). Planned pairwise comparisons between the three word types showed that DAO4 differed significantly from DU4 ($p < .01$), while DEI4 did not differ significantly from DAO4 ($p = .13$) and from DU4 ($p = .24$).

Consistent with a visual inspection of **Figure 5** we found an interaction between *word type* and *step*. Fitting models and carrying out pairwise comparisons for each step separately showed that the DAO4 and DU4 differed significantly for all steps ($p < .01$) except for Steps 1 and 2 ($p = .44$ and $p =.83$, respectively). DAO4 differed from DEI4 significantly for Steps 3 to 7 ($p < .05$) and marginally for Steps 1 and 8 ($p = .08$ and $p = .06$, respectively). Nevertheless, for Steps 6, 7, and 8, DU4 also differed significantly from DEI4. In addition, we found marginal differences between these two word types for Steps 4 ($p = .05$) and 5 ($p = .06$).

In sum, our statistical analysis fully confirmed the visual impression of **Figure 5**. Subjects' responses differed between the three word types. The word type DAO4 led to significantly more T1 responses for the first syllable than DU4, while DEI4 resided between these two. We conclude that, in spite of the same acoustic cues being provided to the listeners in the three word type conditions, the perception of tones in disyllabic stimulus tokens is affected by whether the disyllabic stimuli are real words and real syllables or not.

4.3 Tone Identification in Reduced Disyllables: Onset_S2 Removed

A different group of subjects from the previous experiment was recruited for the second experiment on the reduced stimuli. 36 subjects, 18 female and 18 male, aged between 20 and 30, were paid for the experiment. We removed /t/ in all three continua of *dao4*, *du4*, and *dei4*, including the transition phase to the Nucleus_S2. In total, 72 trials (3 word types x 8 steps x 3 repetitions) were randomized and presented to the subjects using E-Prime. The task of the subjects was to answer whether the first syllable was T1 or T4. The result of the reduced stimuli was completely different from the previous one.

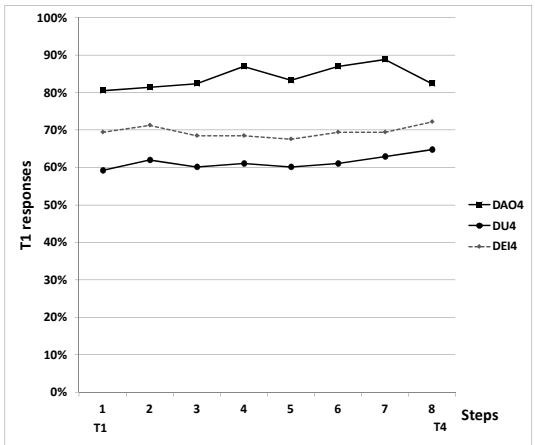

Figure 6. T Tone identification responses (T1) for T1 to T4 continua in reduced disyllables zhi1dao4, zhi1du4, and zhi1dei4.

Visual inspection of **Figure 6** suggests that reduced stimuli had a different effect on tone recognition from the fully pronounced stimuli of the previous experiment. Most obvious, there seems to be no difference in the judgment for the different steps. The listeners constantly tended to hear T1, but with different degrees dependent on word type. When the first syllable was combined with DAO4, which formed a real word, they tended to hear T1 the most. When it was combined with DU4, which formed a nonword, they tended to hear T1 the least. The concatenation with the non-lexical syllable DEI4 lies right in the middle. The question is if the visible differences between the three word types are also statistically significant.

We used the same statistical procedure. Logistic regression models were fit on our data, with *word type* as the fixed factor, and *subjects*, *steps*, *repetitions* as random factors. This time, exclusion of both *steps* and *repetitions* did not lead to a significant decrease of model fit ($p = .99$). This confirms that *step* did not have a significant influence on tone judgment. Nevertheless, adding a *word type* versus *subjects* interaction yielded a significantly better fit ($p < .01$), suggesting that the subjects responded quite differently to the word types. After establishing the random effects structure, we tested for significance of the fixed factor *word type*. The main effect of word type failed to reach significance ($p = 0.13$ in case of DAO4 versus DU4).

As pointed out before, our fitted model contained a significant *word type* versus *subjects* interaction. This suggested that, for some subjects, the difference visible in **Figure 6** between the word types reached significance. We carried out a per-subject analysis. Out of all 36 subjects, we obtained a *word type* effect for 5 subjects. For Subjects 15 and 22, there was a significant difference between DU4 and DEI4 ($p < .01$). For Subject 19, both the differences between DU4 and DEI4 ($p < .05$) and between DU4 and DAO4 ($p < .01$) were significant.

Subject 25 showed significant differences in all comparisons (all $p < .05$).

Overall, the results suggest that when the segmental information is not sufficient, listeners are more likely to interpret the F0 contour as a tone in a disyllabic word in the way that is consistent with an existing word in the mental lexicon. Interestingly, the differences in F0 contour between the steps seem to have been completely overridden by lexical information in this experiment. One caveat to this interpretation is that differences in word type were not statistically significant for all subjects. We suggest that this was due to a ceiling effect (many subjects reached 100% decision scores for T1), and we are currently carrying out further experiments to tackle this issue. In sum, we interpret this result with caution as showing the tendency of an influence of the mental lexicon on tone recognition. In this context, we further suggest that the overall tendency to hear T1 could be due to the higher lexical frequency of *zhi1dao4* versus *zhi4du4*.

4.4 Interim Discussion

We constituted a more concrete association of tones with word meaning through use of disyllabic words, instead of monosyllables, as recognizing tones of monosyllables would not necessarily imply recognizing spoken words. It could also only mean a perceptual processing of pitch differences (Cutler & Chen, 1997; Ye & Connine, 1999). As shown in previous studies, lexical tones can be determined very quickly in monosyllabic words by native listeners (Lai & Zhang, 2008), while non-native listeners also performed quite well in recognizing tones in monosyllabic items (Fox & Unkefer, 1985; Hallé *et al*., 2004). Different from the results proposed by Fox & Unkefer (1985), we did find difference among the tone continua of word-nonword, nonword-word, and nonword-nonword pairs via the fully pronounced disyllabic stimuli. At what stage tones start to function in recognizing spoken words has not yet been studied well. One of the reasons may be that most of the studies have focused on tones of monosyllables produced in isolation. So far, psycholinguistic research is still far from being able to propose a theory of the role of tone in the classical models of spoken word processing, as it is difficult to find a place to include tonal information on the basis of a hierarchy of feature, phoneme, and word (Marslen-Wilson, 1987; McClelland & Elman, 1986). In our results, the non-lexical counterpart *dei4* provides a clear threshold for the distinction between word and nonword from Step 3. This means the lexical effect is observed by shifting the highest F0 value by 20 ms leftwards for *dao4* and *du4* contrasts, which is a very subtle difference. This threshold, however, does not necessarily tell us how tones are processed. As previous studies have pointed out, tonal information in disyllabic words varies to a large extent depending on the tonal context (Xu, 1997) and the perception of tones in multiple syllables are dependent on the pitch register and duration of the following syllable (Lin & Wang, 1984). Thus, tone processing is complex. Could it be possible that tonal

information is processed in a way similar to the late top-down contextual information, as mentioned in the Cohort Theory (Marslen-Wilson, 1987), but maybe at a different level? The recognition rates of tone continua of monosyllabic stimulus tokens in Fox & Unkefer (1985) and Hallé *et al*. (2004) range from nearly 0% to 100%. In our experiment with the fully pronounced disyllables, however, the rates range from 30% to 90%. This difference suggests that the mapping between tone categories and lexical decision deviates depending on how tones are associated with the reduced word forms stored in the lexicon (or in the memory). Furthermore, based on the results of our first experiment, lexical processing of fully pronounced disyllabic words may involve a direct activation of mapping the phonetic form into the canonical form. When segment information is missing, however, acoustic cues of tones did not play a role at all, as no step difference was found. Instead, the response preference was related to the more frequently heard reduced word forms. This suggests that listeners tend to hear the words (thus, the tone) as those they encounter more often in speech communication. Production frequency of words may play a role.

5. Observations in a Case Study of Reduced T1-T4 Disyllable Recognition

Reduction does not result in any obvious problems hindering the understanding of the listeners if enough contextual information is available to them (Ernestus *et al*., 2002). Nevertheless, the question remains of what kinds of cues are there for language users to use to associate a reduced, ambiguous sound (including segmental and tonal information) with a meaning when there is no contextual information. In this section, we conducted a re-analysis of a previous word identification experiment (Tseng & Lee, 2010). Tokens of 8 disyllabic words and 8 disyllables composed of two monosyllabic words that were spoken in two reduction degrees (contracted and merged) and were extracted from the Taiwan Mandarin Conversational Corpus were used as stimuli. Uncontracted stimuli were recorded separately by the same speaker. They were presented to 48 subjects in three conditions: with the full context, in isolation, and in a carrier sentence. The subjects had to identify what they heard in terms of Chinese characters or the phonetic symbols of Zhuyin Fuhao, which is a phonetic transcription method used widely in Taiwan. As all stimulus tokens were spoken by the same speaker, discrepancies of pitch height and pitch range should not be an obvious problem. The previous results in Tseng & Lee (2010) showed that reduced words in their original sentential context were easier to recognize than those in an irrelevant context, such as a less meaningful carrier sentence or in isolation. It was more difficult to recognize more reduced than less reduced words. In consideration of word unit type, it also showed that semantically coherent disyllabic words were more correctly identified than combinations of monosyllabic words that were semantically ambiguous without context. This effect was especially apparent under the isolation condition. In Tseng & Lee (2010), correct responses were counted and analyzed

without examining all of the responses given by the subjects. In the current analysis, we closely transcribed and classified the responses to two T1-T4 disyllable stimuli. As word unit type played a role in the accuracy rate of word identification, we analyzed both *yin1wei4* /in wei/ (because) and *ta1-shi4* /ta ʂɻ̩/ (it/he/she - to be). *Yin1wei4* is a disyllabic word, and *ta1-shi4* a disyllable with two monosyllabic words.

5.1 Word Identification Responses

For *yin1wei4*, if provided with full context, all 48 subjects had no problems recognizing it in both of the contracted and merged cases, as shown in **Table 4**. All subjects used the correct Chinese characters to transcribe it. Nevertheless, when the context was missing, only 24 subjects recognized *yin1wei4* in the contracted case. When the reduction degree increased, the percentage of the correct answers also declined. In the merged cases, only 17 answers were correct. Among the not-recognized responses, there was a consensus on the onset and nucleus of the first syllable. The recognition of segments did not differ much among the subjects, but that of tones did. Different from *yin1wei4*, even provided with full context, not all subjects correctly recognized *ta1-shi4*. 34 subjects recognized it in the contracted case and only 27 in the merged case. To a certain extent, the degree of reduction and the word boundary within the disyllable led to recognition problems. Nevertheless, regardless of whether full context was provided or not, if the responses were not completely correct, at least the first syllable was correctly recognized by the subjects. Even in the merged case without context, 16 answered with *ta1* and 28 subjects with the correct onset of the first syllable. Here, it was shown that the reduction patterns of disyllables differ in terms of the pattern of lexical constituency. Disyllabic words and two monosyllabic words, although both are disyllables, use different reduction devices. The first onset of *ta1-shi4* is mostly preserved, and the second syllable is completely reduced and weakened. Thus, the subjects consistently recognized the first tone as T1. For *yin1wei4*, however, in spite of some consensus on the onset consonant and nucleus, the recognition of tones differed to a large extent. As this may have to do with the F0 shapes of the stimulus tokens, we further analyzed the F0 patterns.

5.2 F0 Shape and Exposure Frequency

We stylized the F0 shapes of the disyllables with two linear lines for the voiced regions in each syllable and calculated the slopes. Baseline slopes were obtained by using the first syllable of the uncontracted tokens that were recorded by the same female speaker. If the slopes of the stylized F0 contour in the contracted and merged tokens rose more than the rising T2, we regarded the F0 shape as a “rise”. If the slopes fell more than the falling T4, we regarded it as a “fall”. Otherwise, the F0 contour was classified as “level”. The F0 shape was categorized only if the duration ratio of the stretch of the F0 was larger than one-third of the

entire syllable, as it was not representative enough otherwise. Applying this procedure, the F0 shape of the contracted *yin1wei4* is rise-fall. The merged *yin1wei4* has a level F0 shape. The contracted stimulus tokens of *ta1-shi4* both have a level F0 shape.

Table 4. Word identification responses to reduced disyllables.

Yin1wei4	Contracted: [inei], rise-fall		Merged: [ỹɛ̃], level	
With context	**Yin1wei4**	48	**Yin1wei4**	48
In isolation	**Yin1wei4**	24	**Yin1wei4**	17
	Onset_S1 recognized as the nasal /n/ or /m/ (S1 answered with T1: 3, T2: 6, T3: 5, T4: 2, T5: 1)	17	Onset_S1 recognized as the bilabial stop /p/ (S1 answered with T1: 2, T4: 6)	8
	Onset_S1 recognized as zero onset with /i/ or /y/ (S1 answered with T1: 6, T2: 1)	7	Onset_S1 recognized as zero onset with /i/ or /y/ (S1 answered with T1: 3, T4: 10, T5: 5)	18
			Others	5
Ta1-shi4	Contracted: [t^haz̩], level		Merged: [t^haʔ], level	
With context	**Ta1-shi4**	34	**Ta1-shi4**	27
	Ta1-de5	13	**Ta1/ta1-zai4/ta1-jiu4**	16
	Not answered	1	Others	5
In isolation	**Ta1-shi4**	9	**Ta1**	16
	Ta1-de5	30	Onset_S1 recognized as /t^h/ (S1 answered with T1: 19)	28
	Ta1-others	6		
	Others	3	Others	4

As shown in **Table 4**, more answers were correct in the recognition of the citation tone of *ta1-shi4* (45 in contracted cases and 35 in merged cases) than of *yin1wei4* (33 in contracted cases and 22 in merged cases). This may be due to the reason that the F0 shapes of *ta1-shi4* are level. Still, the reduced disyllables were easier to recognize if the disyllables formed a word unit. When the contracted *yin1wei4* was spoken in a rise-fall F0 shape, it clearly indicated that there were two syllables. This was correctly recognized by 24 subjects. The merged *yin1wei4* token was produced with a level F0 shape, but 17 subjects were still able to recognize it. *Yin1wei4* is a frequently produced item, so its reduced phonetic form should be familiar to the subjects. In contrast, the internal word boundary in *ta1-shi4* and the greatly reduced second syllable seemed to hinder the recognition; only nine answers were correct in

the contracted case and none in the merged case.

We also examined the exposure frequency of *yin1wei4* and *ta1-shi4* by calculating the bi-gram frequency from the Taiwan Mandarin Conversational Corpus. *Ta1-shi4* appeared 376 times, and *yin1wei4* appeared 1,573 times. The disyllable *yin1wei4* empirically is encountered more often in conversational speech than *ta1-shi4*. The other responses *ta1-de5* and *ta1-jiu4* occur 406 and 259 times, respectively. When segmental information is largely missing, tone information does not seem to act the same way as when the segmental information is complete. The mapping to a lexical representation via a phonetic form encountered often could be more efficient than via a sequential or hierarchical mapping route through the abstract units. As shown in **Table 4**, with insufficient segmental information, the identification of tone categories seems to be closely bound to the "already" recognized words, instead of being processed separately. Exposure frequency may affect the method of mapping a reduced, ambiguous phonetic form to a meaningful word, as it normally determines the degree of familiarity of the "ambiguous phonetic forms".

6. Conclusion

The recognition process of tone and meaning of spoken words may differ when listeners are conducting tasks with different extents of linguistic information, as proposed by Ye & Connine (1999). When presented in isolation, surface phonetic forms of spoken words that are often heard in everyday speech communication influence the way in which we recognize reduced words. This was shown in our tone category identification experiment with reduced stimulus tokens. The simulated disyllabic tokens became ambiguous when the onset consonant of the second syllable was removed, as a direct association of the phonetic form with a canonical word form was nearly impossible. At the same time, due to a clearly audible boundary between the two syllables, we can be sure that the subjects should map the tokens they heard with disyllables. The preference for tones towards real words disappeared with the ambiguous reduced stimuli. Distinction in the acoustic information had no effect on the decision of tone selection, as was evident from the absence of a step effect. We would like to interpret this result as a kind of support for the influence from the production frequency of spoken words (Jurafsky *et al.*, 2001; Myers & Li, 2009). With insufficient segment information, a word with a high production frequency would attract responses towards the word. In our case, as *zhi1dao4* is a frequently spoken word in conversation, its reduced form is more familiar than *zhi4du4*. As a consequence, the preference for hearing a T1 could simply be a frequency effect. Listeners tended to base their judgment on real words, thus more T1 for *zhi1dao4*, less T1 for *zhi4du4*, and the non-lexical counterpart with *dei4* sat in between. Moreover, the word identification experiment suggested that the segmental and tonal information of reduced, ambiguous word forms were recognized as one unit, not separately.

The recognition of spoken words would be based mainly on how frequent the surface phonetic forms, *i.e*., reduced word forms, language users encounter and store/memorize in their mental lexicon. In the consideration of word unit type, production frequency should not merely be lexical frequency, but we should also take into account co-occurrences of words (n-grams). Is production frequency, however, the only factor that determines the recognition of reduced words and the way words are stored in the mental lexicon? To answer this question, we plan to use words that are used rarely in the spoken language as stimuli to study the effect of production frequency. Finally, with regard to our experimental design, it is possible that listeners may shift to different strategies when encountering clear versus reduced speech (see the word type versus subject interaction in the experiment with reduced stimuli). At this point, we cannot determine whether this shift is automatic and rapid (as it should be in a natural conversation that includes both clear and reduced speech) or is induced by the blocked design of our (and other researchers') experiments. We plan to include both clear and reduced stimuli in our next experiments and to further explore the role of individual strategies in listening to natural speech.

Acknowledgements

The authors are grateful to the useful comments provided by two anonymous reviewers of the *International Journal of Computational Linguistics and Chinese Language Processing*. The authors also sincerely thank the team members who have been working on the corpus data along the years. The studies presented in this article were funded by Academia Sinica and the National Science Council under Grant NSC-101-2410-H-001-085.

References

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bertrand, R., Blache, P., Espesser, R., Ferre', G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID: Corpus of Interactional Data. Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, *49*(3), 1-30.

Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.48, retrieved 1 May 2013 from http://www.praat.org/.

Chen, K.J. & Huang, C.-R. (1996). Sinica corpus: Design methodology for balanced corpora. *PACLIC 11*, 167-176.

Cheng, R. L. (1985). Sub-Syllabic Morphemes in Taiwanese. *Journal of Chinese Linguistics*, *13*(1), 12-43.

Chung, R. (1997). Syllable Contraction in Chinese. In *Chinese Languages and Linguistics III. Morphology and Lexicon*. Tsao and Wang (Eds.). Symposium Series of the Institute of History and Philology. Academia Sinica. Taipei. 199-235.

CKIP. (1998). *The Sinica Corpus 3.0*. The Chinese Knowledge Information Processing Group - technical report 98-04. Academia Sinica. (In Chinese)

Cutler, A. & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Attention, Perception, & Psychophysics, 59*(2), 165-179.

Duanmu, S. (2007). *The Phonology of Standard Chinese*. 2nd Edition. Oxford University Press.

Ernestus, M., Baayan, H., & Schreuder, R. (2002). The Recognition of Reduced Word Forms. *Brain and Language*, *81*(1-3), 162-173.

Fox, R. A. & Unkefer, J. (1985). The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics*, *13*(2), 69-89.

Frauenfelder, U.H. & Tyler, L.K. (1987). The process of spoken word recognition: An introduction. *Cognition*, *25*(1-2), 1-20.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110-125.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166-83.

Goddijn, S.M.A. & Binnenpoorte, D. (2003). Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 1361-1364.

Grant, K.W. & Seitz, P.F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *The Journal of the Acoustical Society of America*, *107*, 1000-1011.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267-283.

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics, 32*(3), 395-421.

Ho, D. (1996). *Concept and Method of Phonology*. Daan Publishing (in Chinese).

Hsu, H. (2003). A Sonority Model of Syllable Contraction in Taiwanese Southern Min. *Journal of East Asian Linguistics*, *12*, 349-377.

IPDS. (1995). The Kiel Corpus of Spontaneous Speech. IPDS vol. 1. University of Kiel.

Jeffreys, H. (1961). *The Theory of Probability* (3 ed.). Oxford: Oxford University Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, D. (2001). Probabilistic relations between words: evidence from reduction in lexical production. In Bybee, J. and Hopper, P. (Eds.)

Frequency and the emergence of linguistic structure, 229-54. Amsterdam: John Benjamins.

Kuo, J.-W., Lo, H.-Y., & Wang, H.-M. (2007). Improved HMM/SVM methods for automatic phoneme segmentation. *Interspeech 2007*, Antwerp, Belgium. 2057-2060.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In Elenius, K. & Branderyd, P. (Eds.), *The XIIIth International Congress of Phonetic Sciences.* Stockholm, Sweden. 140-147.

Lai, Y. & Zhang, J. (2008). Mandarin lexical tone recognition: The gating paradigm. *Kansas Working Papers in Linguistics*, *30*, 183-194.

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

Lee, C. Y., Tao, L., & Bond, Z. S. (2008). Identification of acoustically modified Mandarin tones by native listeners. *Journal of Phonetics, 36*(4), 537-563.

Li, A., Zheng, F., & Byrne, W. (2000). CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 485-488.

Lien, C. (1997). Studies on Directional Complement in Taiwan Southern Min – Dialect Typology and Historical Study. In Chiu-yu Tseng (Ed.), *Chinese Languages and Linguistics IV, Typological Studies of Languages in China*, vol. 2. Academia Sinica, Taipei. 379-404.

Lin, T. & Wang, W. S.-Y. (1984). On the issues of tone perception. *Zhongguo Yuwen Xuebao*, *2*, 59-69. (in Chinese)

Liu, Y.-F., Tseng, S.-C., & Chang, R. J.-S. (2013). A phone-aligned conversational corpus of Taiwan Mandarin. Manuscript.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, *49*, 606-608.

Luce, P. A. & D. B. Pisoni. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1-36.

Lung, Y. (1979). A Discussion of the Theory that Yin-sheng Words End with Final Consonants. *Bulletin of the Institute of History and Philology*. *50*(4), 679-716. (in Chinese)

Maekawa, K. (2009). Analysis of language variation using a large-scale corpus of spontaneous speech. In S.-C. Tseng (Ed.) *Linguistic Patterns in Spontaneous Speech*, 27-50. Academia Sinica: Taipei.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71-102.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Meunier, C. & Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, *39*(3), 271-278.

Myers, J. & Li, Y. (2009). Lexical frequency effect in Taiwan Southern Min syllable contraction. *Journal of Phonetics*, *37*, 212-230.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2006). Buckeye Corpus of Conversational Speech (1st release). Columbus, OH: Department of Psychology, Ohio State University, USA.

Ranbom, L. J. & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, *57*, 273-298.

Scharenborg, O., Wan, V., & Ernestus, M. (2010). Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *Journal of the Acoustical Society of America*, *127*(2), 1084-1095.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools, Inc.

Tseng, S.-C. (2005). Syllable Contractions in a Mandarin Conversational Dialogue Corpus. *International Journal of Corpus Linguistics*, *10*(1), 63-83.

Tseng, S.-C. (2013). Lexical coverage in Taiwan Mandarin conversation. *International Journal for Computational Linguistics and Chinese Language Processing*, *18*(1), 1-18.

Tseng, S.-C. & Lee, T.-L. (2010). Contextual effects in recognizing reduced words in spontaneous speech. *Proceedings of DiSS-LPSS Joint Workshop*, Tokyo. 39-42.

Wei, P.-c., Thompson, P. M., Liu, C.-h., Huang, C.-R., & Sun, C. (1997). Historical corpora for synchronic and diachronic linguistics studies. *International Journal for Computational Linguistics and Chinese Language Processing*, *2*(1), 131-145.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25,* 61-84.

Xu, Y. (2004). Understanding tone from the perspective of production and perception. *Language and Linguistics, 5*(4), 757-797.

Ye, Y. & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes*, *14*(5-6), 609-630.

Appendix A

List of Selected Disyllabic Nouns and Verbs

Word	Tokens	Word	Tokens	Word	Tokens
shi2hou4(time)	226	kai1shi3(start)	21	xiao3de2(know)	7
jue2de2(think)	216	jing1ji4(economy)	21	xiang3yao4(want)	7
xian4zai4(now)	165	bie2ren2(someone else)	20	dian4hua4(telephone)	6
zi4ji3(oneself)	82	tong2xue2(classmate)	18	dian4ying3(movie)	4
zhi1dao4(know)	80	xiao3hai2(children)	15	zi4ran2(nature)	4
dong1xi1(things)	60	ping2chang2(usually)	17	gang1cai2(just)	4
da4jia1(all)	47	ji4de2(remember)	17	fa1sheng1(happen)	4
hou4lai2(later)	45	xu1yao4(need)	16	sheng1yi4(business)	3
jin1tian1(today)	44	zui4hou4(at last)	16	sheng1huo2(life)	3
jie2guo3(result)	42	guo2wang2(king)	16	guo2yu3	2
ban4fa3(solution)	29	xi1wang4(hope)	15	(national language)	
yuan4yi4(will)	27	deng3yu2(equal)	13	shu3yu2(belong)	2
xi3huan1(like)	26	she4hui4(society)	12	xia4wu3(afternoon)	2
bian4cheng2(turn into)	26	zheng4fu3(government)	12	yao4shi4(what if)	1
jie2yun4(MRT)	25	dian4shi4(TV)	11	nyu3hai2(girls)	1
peng2you3(friend)	24	shi4shi2(fact)	10	di4li3(geography)	1
gan3jue2(feel)	22	xian1sheng1(Mr.)	8	jia1ren2(family)	1